

Comparación del desempeño de modelos de aprendizaje y métodos de fusión con técnicas de reducción de características en los datasets MHealth, UTD y PMAP2.

Basto, Manuel Calderón, Carlos Chacón, David Quevedo, Jonathan

Ramírez, Cristhian

08 de Mayo del 2025

Resumen

El reconocimiento de actividades humanas (HAR) mediante sensores iniciales y fisiológicos ha adquirido gran relevancia en aplicaciones biomédicas, deportivas y de interacción humano-computadora. Este trabajo presenta una evaluación comparativa del desempeño de múltiples modelos de aprendizaje automático (Random Forest, Perceptron, Regresión Logística, MLP y SVM) combinados con diversas técnicas de reducción de características (PCA, Kernel PCA, ANOVA F-test) y métodos de fusión (Voting y Boosting), aplicados a tres datasets de referencia: MHealth, UTD-MHAD y PAMAP2. A través de un enfoque experimental, se analiza el impacto de cada técnica sobre métricas clave como accuracy y F1-score. Los resultados evidencian que Random Forest, junto con reducción de características mediante ANOVA F-test, ofrece el mejor equilibrio entre rendimiento y eficiencia computacional.

versas acciones humanas [16][33][13]. Sin embargo, estos datos suelen ser de alta dimensionalidad, ruidosos y heterogéneos, lo que plantea desafíos importantes para el diseño de modelos predictivos eficientes y precisos [14][11].

Este estudio tiene como objetivo comparar el desempeño de distintos modelos de aprendizaje automático (Random Forest, Perceptron, Regresión Logística, MLP y SVM) y métodos de fusión (Voting, Boosting, Random Forest como agregador) aplicados sobre tres conjuntos de datos ampliamente utilizados en HAR: MHealth [36], UTD-MHAD [16], y PAMAP2 [37]. Además, se evalúa el impacto de diversas técnicas de reducción de dimensionalidad (PCA [29], Kernel PCA [34], y ANOVA F-test [38]) sobre la precisión y capacidad de generalización de los modelos.

1. Introducción

El reconocimiento de actividades humanas (HAR, por sus siglas en inglés) es un área de investigación que ha cobrado gran relevancia en aplicaciones como la salud móvil, la rehabilitación, el monitoreo deportivo y la interacción humano-computadora. En este contexto, los datos multi-variados capturados por sensores iniciales y fisiológicos —como acelerómetros, giroscopios, magnetómetros y electrocardiogramas— permiten caracterizar patrones de movimiento asociados a di-

A través de una serie de experimentos sistemáticos, este trabajo busca identificar las combinaciones de modelos, técnicas de fusión y estrategias de reducción de características que ofrecen el mejor rendimiento en tareas de clasificación multiclas con datos sensoriales complejos. El análisis considera tanto métricas tradicionales como accuracy y F1-score, como también la estabilidad y escalabilidad de cada enfoque [11][13][28]. Los hallazgos de esta investigación tienen implicaciones prácticas para el diseño de sistemas HAR robustos, capaces de operar eficientemente en contextos del mundo real [28][39][41].

2. Dataset mHealth

2.1. Descripción del dataset

El **mHealth dataset** contiene registros de movimiento corporal y signos vitales de diez voluntarios mientras realizaban 12 actividades físicas distintas. Los datos fueron recopilados utilizando sensores portátiles **Shimmer2**, que se colocaron en el **pecho, muñeca derecha y tobillo izquierdo** de cada sujeto mediante correas elásticas.

2.1.1. Configuración del experimento

Los sensores registraron información sobre:

- Aceleración (acelerómetro)
- Velocidad angular (giroscopio)
- Campo magnético (magnetómetro)
- Electrocardiograma (ECG) (solo desde el sensor del pecho)

Las mediciones se realizaron a una frecuencia de **50 Hz** (50 muestras por segundo), considerada suficiente para capturar la actividad humana. Además, las sesiones fueron grabadas en video para mayor referencia.

Las actividades fueron realizadas en un entorno **fuera del laboratorio**, sin restricciones en la ejecución, excepto la indicación de hacerlas lo mejor posible.

2.1.2. Actividades registradas

Cada actividad tiene una duración o un número específico de repeticiones:

1. De pie sin moverse (1 min)
2. Sentado y relajado (1 min)
3. Acostado (1 min)
4. Caminando (1 min)
5. Subiendo escaleras (1 min)
6. Flexión de cintura hacia adelante (20 repeticiones)
7. Elevación frontal de brazos (20 repeticiones)
8. Flexión de rodillas (en cuclillas) (20 repeticiones)
9. Ciclismo (1 min)
10. Trotando (1 min)
11. Corriendo (1 min)
12. Saltos hacia adelante y atrás (20 repeticiones)

2.1.3. Archivos del dataset

Los datos de cada sujeto se almacenan en un archivo diferente con el formato:

mHealth_subject<SUBJECT_ID>.log

Cada archivo contiene filas con muestras de datos y columnas con mediciones de los sensores. La etiqueta de actividad está en la última columna.

2.1.4. Variables registradas

Las columnas del dataset incluyen:

- Aceleración (X, Y, Z) del pecho, tobillo izquierdo y brazo derecho.
- ECG (dos derivaciones desde el pecho).
- Giroscopio (X, Y, Z) en el tobillo izquierdo y brazo derecho.
- Magnetómetro (X, Y, Z) en el tobillo izquierdo y brazo derecho.
- Etiqueta de actividad (de 1 a 12, con 0 para datos sin actividad).

Las unidades son:

- Aceleración: m/s²
- Giroscopio: grados/s
- Campo magnético: magnitud local
- ECG: mV

Este dataset es útil para el reconocimiento de actividades humanas y el análisis del impacto del ejercicio en la frecuencia cardíaca.

2.2. Carga del dataset

Para este proceso, primero se descargó el conjunto de datos desde el repositorio UCI Irvine, de modo que se genere una carpeta con los siguientes archivos:

■ mHealth_subject1	17/02/2023 14:54 p. m.	Documento de tex.	28,317 kB
■ mHealth_subject2	17/02/2023 14:54 p. m.	Documento de tex.	23,874 kB
■ mHealth_subject3	17/02/2023 14:54 p. m.	Documento de tex.	22,255 kB
■ mHealth_subject4	17/02/2023 14:54 p. m.	Documento de tex.	21,321 kB
■ mHealth_subject5	17/02/2023 14:54 p. m.	Documento de tex.	21,763 kB
■ mHealth_subject6	17/02/2023 14:54 p. m.	Documento de tex.	19,876 kB
■ mHealth_subject7	17/02/2023 14:54 p. m.	Documento de tex.	20,000 kB
■ mHealth_subject8	17/02/2023 14:54 p. m.	Documento de tex.	22,533 kB
■ mHealth_subject9	17/02/2023 14:54 p. m.	Documento de tex.	24,589 kB
■ mHealth_subject10	17/02/2023 14:54 p. m.	Documento de tex.	18,007 kB
■ README	17/02/2023 14:54 p. m.	Documento de tex.	6 kB

Figura 1: Listado de archivos del dataset

Posteriormente, dentro del entorno de Jupyter con Python 3 se procedió a cargar cada archivo .log dentro de un mismo objeto DataFrame de la librería Pandas para facilitar su análisis y manejo:

```
[1]: # Importación de librerías
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import os

[4]: # Año a la carpeta con los archivos del dataset
data_path = "C:/Users/C-Mateus/Documents/LCC/RAMBFOP/ProyectoFinal/PHEALTHDATASET"

# Nombres de las columnas
column_names = ["subject","acc_chest_x","acc_chest_y","acc_chest_z","electrocardiogram_11",
"acc_chest_x_11","acc_chest_y_11","acc_chest_z_11","gyro_left_ankle_x",
"gyro_left_ankle_y","gyro_left_ankle_z","gyro_left_ankle_x_1",
"gyro_left_ankle_y_1","gyro_left_ankle_z_1","MAG_left_ankle_x",
"gyro_r_lw_arm_x","gyro_r_lw_arm_y","gyro_r_lw_arm_z",
"gyro_r_lw_arm_x_1","gyro_r_lw_arm_y_1","gyro_r_lw_arm_z_1",
"label"]

[5]: # Creación de dataframe vacío
data = pd.DataFrame(columns=column_names)

# Carga de datasets de cada sujeto y concatenación a dataframe
for i, file in enumerate(os.listdir(data_path)):
    if file == "README.txt":
        continue
    df_tmp = pd.read_csv(os.path.join(data_path, file), header=None, sep="\t")
    df_tmp.columns = column_names[1:]
    df_tmp["subject"] = i+1
    data = pd.concat([data, df_tmp])


```

Figura 2: Creación de DataFrame para análisis.

2.3. Dimensiones del Dataset y primeras 5 líneas

Siempre es una buena práctica conocer el tamaño de nuestro dataset en términos de filas y columnas, para lograr esto, se hizo uso del atributo shape de nuestro dataframe recién creado.

```
[6]: data.shape
[6]: (1215745, 25)
```

Figura 3: Comando para acceder a las dimensiones del dataset.

De esta manera notamos que nuestro dataframe cuenta con 1215745 filas y 25 columnas.

Asimismo, para darle el primer vistazo a nuestros datos crudos se hizo uso del método head() para desplegar las primeras 5 filas del dataframe, las cuales sirven de ejemplo para darse una idea de la organización actual de los datos.

```
[11]: print(data.head())
   subject acc_chest_x acc_chest_y acc_chest_z electrocardiogram_11 \
0      1     -9.8184  0.089971  0.259683  0.004186
1      1     -9.8489  0.524840  0.37348  0.004186
2      1     -9.8489  0.524840  0.37348  0.004186
3      1     -9.6587  0.212420  0.24083  0.079548
4      1     -9.7089  0.303898  0.31156  0.221870

  electrocardiogram_12 acc_left_ankle_x acc_left_ankle_y acc_left_ankle_z \
0     0.084186  2.1849  -9.4967  0.63077
1     0.084186  2.1876  -9.5808  0.68391
2     0.084186  2.1866  -9.5576  0.62139
3     0.117220  2.1814  -9.4381  0.55631
4     0.205130  2.1473  -9.3889  0.71089

  gyro_left_ankle_x ... acc_r_lw_arm_x acc_r_lw_arm_y acc_r_lw_arm_z \
0     0.109988 ... -8.4599  -8.5781  0.187768
1     0.052570 ... -8.4599  -8.5781  0.187768
2     0.085343 ... -8.3985  -8.2772  0.257528
3     0.085343 ... -8.6279  -8.3163  0.367528
4     0.085343 ... -8.7908  -8.1459  0.407288

  gyro_r_lw_arm_x gyro_r_lw_arm_y gyro_r_lw_arm_z MAG_r_lw_arm_x \
0     -0.44902  0.010483  0.010483  -2.358088
1     -0.44902  0.034483  0.034483  -2.358088
2     -0.44902  -1.0183  0.034483  -1.61758
3     -0.45668  -1.0082  0.025862  -1.07718
4     -0.45668  -1.0082  0.025862  -0.554584

  MAG_r_lw_arm_y MAG_r_lw_arm_z label
0      0.118802  0.023582  0
1     -0.882549  0.126579  0
2     -0.165628  -0.036093  0
3     0.086945  -0.382629  0
4     0.175988  -1.495598  0
```

Figura 4: Despliegue de primeras 5 filas del dataset.

2.4. Datos Ausentes

El siguiente paso que se realizó fue la identificación de los datos ausentes/nulos en el dataset. Para esto se hizo uso del método isnull() del dataframe, en conjunto con el método de agregación sum(), combinación que nos permitió obtener un conteo de valores nulos por columna en el dataframe.

```
[10]: data.isnull().sum() # Conteo de nulos por columna
[10]: subject          0
       acc_chest_x       0
       acc_chest_y       0
       acc_chest_z       0
       electrocardiogram_11 0
       electrocardiogram_12 0
       acc_left_ankle_x       0
       acc_left_ankle_y       0
       acc_left_ankle_z       0
       gyro_left_ankle_x       0
       gyro_left_ankle_y       0
       gyro_left_ankle_z       0
       MAG_left_ankle_x       0
       MAG_left_ankle_y       0
       MAG_left_ankle_z       0
       acc_r_lw_arm_x       0
       acc_r_lw_arm_y       0
       acc_r_lw_arm_z       0
       gyro_r_lw_arm_x       0
       gyro_r_lw_arm_y       0
       gyro_r_lw_arm_z       0
       MAG_r_lw_arm_x       0
       MAG_r_lw_arm_y       0
       MAG_r_lw_arm_z       0
       label                0
       dtype: int64
```

Figura 5: Conteo de datos nulos por columna.

Así, al no contar este dataset con la presencia de datos ausentes, no fue necesario realizar algún proceso extra para el manejo de estos.

2.5. Tipos de datos por atributo

Posteriormente, se dio un vistazo a los tipos de datos relacionados a cada atributo/columna del dataset, para esto fue usado el atributo df.dtypes.

```
[12]: data.dtypes # Desplegar tipos de datos de cada atributo
[12]: subject          object
       acc_chest_x      float64
       acc_chest_y      float64
       acc_chest_z      float64
       electrocardiogram_11 float64
       electrocardiogram_12 float64
       acc_left_ankle_x      float64
       acc_left_ankle_y      float64
       acc_left_ankle_z      float64
       gyro_left_ankle_x      float64
       gyro_left_ankle_y      float64
       gyro_left_ankle_z      float64
       MAG_left_ankle_x      float64
       MAG_left_ankle_y      float64
       MAG_left_ankle_z      float64
       acc_r_lw_arm_x      float64
       acc_r_lw_arm_y      float64
       acc_r_lw_arm_z      float64
       gyro_r_lw_arm_x      float64
       gyro_r_lw_arm_y      float64
       gyro_r_lw_arm_z      float64
       MAG_r_lw_arm_x      float64
       MAG_r_lw_arm_y      float64
       MAG_r_lw_arm_z      float64
       label                object
       dtype: object
```

Figura 6: Tipos de dato por atributo.

2.6. Resumen estadístico por atributo

Para esta parte del proceso fueron generados dos resúmenes, un resumen general y otro agrupando a los datos por sujeto de prueba, haciendo uso de los métodos describe() y groupby().

```
[13]: pd.options.display.float_format = '{:.4f}'.format ## Desplegar resultados con 4 decimales de precisión
print(data.describe()) # Imprime resumen estadístico

acc_chest_x acc_chest_y acc_chest_z acc_left_ankle_x
count 1215745.0000 1215745.0000 1215745.0000
mean -8.5224 -0.2140 -15.6559
std 4.0323 0.9909 3.7422
min -22.4398 -26.1889 -16.4910
25% -9.8848 -1.2729 -2.8673
50% -9.3114 -0.3538 -0.8782
-0.0712
75% -7.0093 0.2988 0.2559
max 19.0940 20.9279 26.1969
0.5149

electrocardiogram_12 acc_left_ankle_x acc_left_ankle_y
count 1215745.0000 1215745.0000 1215745.0000
mean -0.0044 1.4942 -0.0929
std 0.7273 3.8262 4.1713
min -8.6196 -22.1469 -19.6198
25% -0.2554 -0.2146 -0.2120
50% -0.0419 1.3889 -0.6793
75% 0.1507 2.5758 -0.0422
max 8.5191 20.8540 21.1618

acc_left_ankle_z gyro_left_ankle_x gyro_left_ankle_y ...
count 1215745.0000 1215745.0000 1215745.0000
mean -0.9540 -0.0016 -0.6168 ...
std 0.6549 0.8454 0.8452
min -19.3730 -2.1466 -7.7599 ...
25% -2.6494 -0.4360 -0.8180 ...
50% -0.0163 -0.0148 -0.0787 ...
75% 0.1493 0.3469 -0.5601 ...
max 25.9150 60.4640 2.0111

MAG_left_ankle_x MAG_r_hum_r_wx MAG_r_hum_r_wy MAG_r_hum_r_wz
count 1215745.0000 1215745.0000 1215745.0000 1215745.0000
mean -0.3564 -3.7130 -5.0855 2.3939
std 17.6822 4.7636 5.7576 3.8765
min -20.3980 -22.3630 -18.9720 -18.2390
25% -2.4531 -2.4368 -4.6412 0.4246
50% -0.4350 -2.7776 -7.4615 1.9281
75% 1.9769 -1.1397 -2.5339 4.9147
max 272.5600 19.8640 22.1916 25.7410

GYRO_r_hum_r_wx GYRO_r_hum_r_wy GYRO_r_hum_r_wz MAG_r_hum_r_wx
count 1215745.0000 1215745.0000 1215745.0000 1215745.0000
mean -0.2761 -0.4664 0.2666 0.1702
std 0.3277 0.3656 0.3444 0.3652
min -8.3392 -3.5798 -2.6897 -319.0300
25% -0.7059 -0.8973 -0.2371 -0.1919
50% -0.3549 -0.6345 0.3817 0.3626
75% 0.9491 -0.6700 0.7700 1.7523
max 3.3186 1.5565 2.7980 251.1598

MAG_r_hum_r_wy MAG_r_hum_r_wz
count 1215745.0000 1215745.0000
mean 0.7145 -0.3668
std 33.4445 69.5597
min -383.9800 -717.5500
25% -199.9500 -389.0000
50% 0.3522 -0.6730
75% 10.0760 13.1860
max 337.7600 657.1800
```

Figura 7: Resumen estadístico general.

```
[14]: pd.options.display.float_format = '{:.3f}'.format ## Desplegar resultados con 3 decimales de precisión
data.groupby(["subject"]).describe() # Imprime resumen estadístico por sujeto

acc_chest_x acc_chest_y acc_chest_z MAG_r_hum_arm_y
subject count mean std min 25% 50% 75% max count mean std min 25% 50%
1 1612800.0000 -8.738 3.806 -22.103 -9.833 -9.331 -8.021 18.960 1612800.0000 0.186 ... 10.905 134.980 1612800.0000 -0.231 62.555 -997.160 -13.399 -0.170
2 9830400.0000 -8.417 4.094 -22.119 -9.758 -9.156 -6.629 17.616 9830400.0000 0.201 ... 10.153 32.059 9830400.0000 -0.184 60.940 504.800 -15.257 -0.715
3 13056100.0000 -8.83 1.601 -22.295 -9.871 -9.399 -8.278 16.526 13056100.0000 -0.187 ... 9.472 136.220 13056100.0000 -0.041 66.663 -401.630 -14.314 -0.159
4 12211200.0000 -8.332 4.145 -22.438 -9.738 -9.156 -7.264 17.605 12211200.0000 -0.194 ... 12.607 353.250 16763000.0000 -0.052 52.193 615.100 -13.306 -0.361
5 11673600.0000 -7.81 4.278 -22.259 -9.604 -9.073 -6.633 17.004 11673600.0000 -0.190 ... 11.873 337.760 12112000.0000 -0.049 74.193 678.360 -19.639 -0.718
6 11980800.0000 -8.637 4.315 -22.168 -10.006 -9.373 -7.736 17.002 11980800.0000 -0.194 ... 11.873 334.450 11980800.0000 -0.034 74.193 678.360 -19.639 -0.718
7 9830400.0000 -8.633 4.007 -22.321 -9.829 -9.468 -7.912 14.857 9830400.0000 -0.224 ... 7.673 310.420 9830400.0000 -0.187 69.244 -600.820 -14.988 -0.716
8 10444800.0000 -8.440 4.185 -22.266 -9.743 -9.177 15.875 10444800.0000 -0.333 ... 7.628 311.170 10444800.0000 -0.301 59.335 -453.910 -15.596 -0.377
9 12004000.0000 -8.558 3.976 -22.289 -9.698 -9.391 -8.246 17.024 12004000.0000 -0.375 ... 8.272 344.120 12004000.0000 -0.087 72.850 -717.550 -9.780 -0.602
10 13116600.0000 -8.625 4.312 -22.306 -9.835 -9.464 -7.538 19.070 13116600.0000 -0.082 ... 10.358 316.120 13116600.0000 -1.003 87.499 -668.430 -16.811 -0.713

10 rows x 184 columns
```

Figura 8: Resumen estadístico por sujeto.

A primera vista podemos notar que los datos que nos brindan una mayor variabilidad son, en su mayoría, aquellos proporcionados por los acelerómetros y magnetómetros (mayor desviación estándar / std).

De igual forma, basándonos en los mínimos y máximos de cada característica, podemos observar una gran diferencia de escalas en algunos casos, por lo que se abre la posibilidad de tener que

realizar un reescalado de características. Esto para mejorar el funcionamiento del modelo seleccionado, puesto que, poniendo de ejemplo el modelo KNN, en [1] se menciona: "La escala de las variables puede desvirtuar el resultado del modelo KNN. Por ello, el conjunto de datos debe escalarse para que las variables con unidades de medida grandes no tengan más importancia en el cálculo de la similitud entre las observaciones que las que tienen magnitudes menores."

Nótese que por cuestiones de formato, no fue posible mostrar la salida completa en este reporte.

2.7. Distribución de clases

Luego, para esta sección también se procedió a mostrar la distribución de clases general y por sujeto de prueba.

Se utilizó el método groupby("label") para agrupar por clase para la primer distribución y groupby(["subject","label"]) para agrupar por sujeto y clase, luego se usó el método size() para el conteo de frecuencias.

```
[30]: class_freq = data.groupby("label").size() # Agrupar por Label y desplegar conteo
class_freq

[30]: label
0 872550
1 30720
2 30720
3 30720
4 30720
5 30720
6 20312
7 29441
8 29337
9 30720
10 30720
11 30720
12 10342
dtype: int64
```

(a) Distribución General.

```
[31]: class_freq_p_subject = data.groupby(["subject","label"]).size() # Agrupar por sujeto y Label y desplegar conteo
class_freq_p_subject

[31]: subject label
1 0 130400
1 1 9872
1 2 3072
1 3 3072
1 4 3072
1 5 3072
1 6 3072
1 7 3072
1 8 3072
1 9 3072
1 10 3072
1 11 3072
1 12 1024
1 13 1024
1 14 64334
1 15 3072
1 16 3072
1 17 3072
1 18 3072
1 19 3072
1 20 3072
1 21 3072
1 22 3072
1 23 3072
1 24 3072
1 25 3072
1 26 3072
1 27 3072
1 28 3072
1 29 3072
1 30 3072
1 31 3072
1 32 3072
1 33 3072
1 34 3072
1 35 3072
1 36 3072
1 37 3072
1 38 3072
1 39 3072
1 40 3072
1 41 3072
1 42 3072
1 43 3072
1 44 3072
1 45 3072
1 46 3072
1 47 3072
1 48 3072
1 49 3072
1 50 3072
1 51 3072
1 52 3072
1 53 3072
1 54 3072
1 55 3072
1 56 3072
1 57 3072
1 58 3072
1 59 3072
1 60 3072
1 61 3072
1 62 3072
1 63 3072
1 64 3072
1 65 3072
1 66 3072
1 67 3072
1 68 3072
1 69 3072
1 70 3072
1 71 3072
1 72 3072
1 73 3072
1 74 3072
1 75 3072
1 76 3072
1 77 3072
1 78 3072
1 79 3072
1 80 3072
1 81 3072
1 82 3072
1 83 3072
1 84 3072
1 85 3072
1 86 3072
1 87 3072
1 88 3072
1 89 3072
1 90 3072
1 91 3072
1 92 3072
1 93 3072
1 94 3072
1 95 3072
1 96 3072
1 97 3072
1 98 3072
1 99 3072
1 100 3072
1 101 3072
1 102 3072
1 103 3072
1 104 3072
1 105 3072
1 106 3072
1 107 3072
1 108 3072
1 109 3072
1 110 3072
1 111 3072
1 112 3072
1 113 3072
1 114 3072
1 115 3072
1 116 3072
1 117 3072
1 118 3072
1 119 3072
1 120 3072
1 121 3072
1 122 3072
1 123 3072
1 124 3072
1 125 3072
1 126 3072
1 127 3072
1 128 3072
1 129 3072
1 130 3072
1 131 3072
1 132 3072
1 133 3072
1 134 3072
1 135 3072
1 136 3072
1 137 3072
1 138 3072
1 139 3072
1 140 3072
1 141 3072
1 142 3072
1 143 3072
1 144 3072
1 145 3072
1 146 3072
1 147 3072
1 148 3072
1 149 3072
1 150 3072
1 151 3072
1 152 3072
1 153 3072
1 154 3072
1 155 3072
1 156 3072
1 157 3072
1 158 3072
1 159 3072
1 160 3072
1 161 3072
1 162 3072
1 163 3072
1 164 3072
1 165 3072
1 166 3072
1 167 3072
1 168 3072
1 169 3072
1 170 3072
1 171 3072
1 172 3072
1 173 3072
1 174 3072
1 175 3072
1 176 3072
1 177 3072
1 178 3072
1 179 3072
1 180 3072
1 181 3072
1 182 3072
1 183 3072
1 184 3072
1 185 3072
1 186 3072
1 187 3072
1 188 3072
1 189 3072
1 190 3072
1 191 3072
1 192 3072
1 193 3072
1 194 3072
1 195 3072
1 196 3072
1 197 3072
1 198 3072
1 199 3072
1 200 3072
1 201 3072
1 202 3072
1 203 3072
1 204 3072
1 205 3072
1 206 3072
1 207 3072
1 208 3072
1 209 3072
1 210 3072
1 211 3072
1 212 3072
1 213 3072
1 214 3072
1 215 3072
1 216 3072
1 217 3072
1 218 3072
1 219 3072
1 220 3072
1 221 3072
1 222 3072
1 223 3072
1 224 3072
1 225 3072
1 226 3072
1 227 3072
1 228 3072
1 229 3072
1 230 3072
1 231 3072
1 232 3072
1 233 3072
1 234 3072
1 235 3072
1 236 3072
1 237 3072
1 238 3072
1 239 3072
1 240 3072
1 241 3072
1 242 3072
1 243 3072
1 244 3072
1 245 3072
1 246 3072
1 247 3072
1 248 3072
1 249 3072
1 250 3072
1 251 3072
1 252 3072
1 253 3072
1 254 3072
1 255 3072
1 256 3072
1 257 3072
1 258 3072
1 259 3072
1 260 3072
1 261 3072
1 262 3072
1 263 3072
1 264 3072
1 265 3072
1 266 3072
1 267 3072
1 268 3072
1 269 3072
1 270 3072
1 271 3072
1 272 3072
1 273 3072
1 274 3072
1 275 3072
1 276 3072
1 277 3072
1 278 3072
1 279 3072
1 280 3072
1 281 3072
1 282 3072
1 283 3072
1 284 3072
1 285 3072
1 286 3072
1 287 3072
1 288 3072
1 289 3072
1 290 3072
1 291 3072
1 292 3072
1 293 3072
1 294 3072
1 295 3072
1 296 3072
1 297 3072
1 298 3072
1 299 3072
1 300 3072
1 301 3072
1 302 3072
1 303 3072
1 304 3072
1 305 3072
1 306 3072
1 307 3072
1 308 3072
1 309 3072
1 310 3072
1 311 3072
1 312 3072
1 313 3072
1 314 3072
1 315 3072
1 316 3072
1 317 3072
1 318 3072
1 319 3072
1 320 3072
1 321 3072
1 322 3072
1 323 3072
1 324 3072
1 325 3072
1 326 3072
1 327 3072
1 328 3072
1 329 3072
1 330 3072
1 331 3072
1 332 3072
1 333 3072
1 334 3072
1 335 3072
1 336 3072
1 337 3072
1 338 3072
1 339 3072
1 340 3072
1 341 3072
1 342 3072
1 343 3072
1 344 3072
1 345 3072
1 346 3072
1 347 3072
1 348 3072
1 349 3072
1 350 3072
1 351 3072
1 352 3072
1 353 3072
1 354 3072
1 355 3072
1 356 3072
1 357 3072
1 358 3072
1 359 3072
1 360 3072
1 361 3072
1 362 3072
1 363 3072
1 364 3072
1 365 3072
1 366 3072
1 367 3072
1 368 3072
1 369 3072
1 370 3072
1 371 3072
1 372 3072
1 373 3072
1 374 3072
1 375 3072
1 376 3072
1 377 3072
1 378 3072
1 379 3072
1 380 3072
1 381 3072
1 382 3072
1 383 3072
1 384 3072
1 385 3072
1 386 3072
1 387 3072
1 388 3072
1 389 3072
1 390 3072
1 391 3072
1 392 3072
1 393 3072
1 394 3072
1 395 3072
1 396 3072
1 397 3072
1 398 3072
1 399 3072
1 400 3072
1 401 3072
1 402 3072
1 403 3072
1 404 3072
1 405 3072
1 406 3072
1 407 3072
1 408 3072
1 409 3072
1 410 3072
1 411 3072
1 412 3072
1 413 3072
1 414 3072
1 415 3072
1 416 3072
1 417 3072
1 418 3072
1 419 3072
1 420 3072
1 421 3072
1 422 3072
1 423 3072
1 424 3072
1 425 3072
1 426 3072
1 427 3072
1 428 3072
1 429 3072
1 430 3072
1 431 3072
1 432 3072
1 433 3072
1 434 3072
1 435 3072
1 436 3072
1 437 3072
1 438 3072
1 439 3072
1 440 3072
1 441 3072
1 442 3072
1 443 3072
1 444 3072
1 445 3072
1 446 3072
1 447 3072
1 448 3072
1 449 3072
1 450 3072
1 451 3072
1 452 3072
1 453 3072
1 454 3072
1 455 3072
1 456 3072
1 457 3072
1 458 3072
1 459 3072
1 460 3072
1 461 3072
1 462 3072
1 463 3072
1 464 3072
1 465 3072
1 466 3072
1 467 3072
1 468 3072
1 469 3072
1 470 3072
1 471 3072
1 472 3072
1 473 3072
1 474 3072
1 475 3072
1 476 3072
1 477 3072
1 478 3072
1 479 3072
1 480 3072
1 481 3072
1 482 3072
1 483 3072
1 484 3072
1 485 3072
1 486 3072
1 487 3072
1 488 3072
1 489 3072
1 490 3072
1 491 3072
1 492 3072
1 493 3072
1 494 3072
1 495 3072
1 496 3072
1 497 3072
1 498 3072
1 499 3072
1 500 3072
1 501 3072
1 502 3072
1 503 3072
1 504 3072
1 505 3072
1 506 3072
1 507 3072
1 508 3072
1 509 3072
1 510 3072
1 511 3072
1 512 3072
1 513 3072
1 514 3072
1 515 3072
1 516 3072
1 517 3072
1 518 3072
1 519 3072
1 520 3072
1 521 3072
1 522 3072
1 523 3072
1 524 3072
1 525 3072
1 526 3072
1 527 3072
1 528 3072
1 529 3072
1 530 3072
1 531 3072
1 532 3072
1 533 3072
1 534 3072
1 535 3072
1 536 3072
1 537 3072
1 538 3072
1 539 3072
1 540 3072
1 541 3072
1 542 3072
1 543 3072
1 544 3072
1 545 3072
1 546 3072
1 547 3072
1 548 3072
1 549 3072
1 550 3072
1 551 3072
1 552 3072
1 553 3072
1 554 3072
1 555 3072
1 556 3072
1 557 3072
1 558 3072
1 559 3072
1 560 3072
1 561 3072
1 562 3072
1 563 3072
1 564 3072
1 565 3072
1 566 3072
1 567 3072
1 568 3072
1 569 3072
1 570 3072
1 571 3072
1 572 3072
1 573 3072
1 574 3072
1 575 3072
1 576 3072
1 577 3072
1 578 3072
1 579 3072
1 580 3072
1 581 3072
1 582 3072
1 583 3072
1 584 3072
1 585 3072
1 586 3072
1 587 3072
1 588 3072
1 589 3072
1 590 3072
1 591 3072
1 592 3072
1 593 3072
1 594 3072
1 595 3072
1 596 3072
1 597 3072
1 598 3072
1 599 3072
1 600 3072
1 601 3072
1 602 3072
1 603 3072
1 604 3072
1 605 3072
1 606 3072
1 607 3072
1 608 3072
1 609 3072
1 610 3072
1 611 3072
1 612 3072
1 613 3072
1 614 3072
1 615 3072
1 616 3072
1 617 3072
1 618 3072
1 619 3072
1 620 3072
1 621 3072
1 622 3072
1 623 3072
1 624 3072
1 625 3072
1 626 3072
1 627 3072
1 628 3072
1 629 3072
1 630 3072
1 631 3072
1 632 3072
1 633 3072
1 634 3072
1 635 3072
1 636 3072
1 637 3072
1 638 3072
1 639 3072
1 640 3072
1 641 3072
1 642 3072
1 643 3072
1 644 3072
1 645 3072
1 646 3072
1 647 3072
1 648 3072
1 649 3072
1 650 3072
1 651 3072
1 652 3072
1 653 3072
1 654 3072
1 655 3072
1 656 3072
1 657 3072
1 658 3072
1 659 3072
1 660 3072
1 661 3072
1 662 3072
```

Viendo nuestra distribución general se pudo notar que el dataset completo está desbalanceado, pues más del 70 % de los datos pertenece la clase 0 (actividad nula), al igual que la actividad 12 (saltos hacia adelante y atrás), es la que menos frecuencia tiene comparada a las demás.

De igual forma, se tuvo que las distribuciones por sujeto son muy similares a la distribución general, pues guardan casi la misma proporción entre clases.

La literatura ([3]) menciona que una distribución desequilibrada de clases en datos de series temporales a menudo resulta en un sesgo en la clasificación de superficies, lo que impide que el clasificador logre el mejor rendimiento. Y cuando el clasificador no puede detectar una clase minoritaria de datos, esto indica que el modelo ha fallado. En el mundo real, la incapacidad de los modelos para clasificar las clases minoritarias de datos (eventos raros) puede tener consecuencias graves.

Por lo anterior, sería necesario profundizar en la necesidad de aplicar alguna técnica de resampling de los datos, para balancear nuestra distribución.

2.8. Correlación entre atributos

Para presentar la correlación entre atributos se hizo uso del método `df.corr()`, como se muestra a continuación:

```
[56]: print(data.corr()) # Desplegamos la correlacion de Pearson entre atributos
```

Figura 10: Comando para obtener la correlación entre atributos.

	subject	acc_chest_x	acc_chest_y	acc_chest_z	\
subject	1.000	0.095	-0.124	0.061	
acc_chest_x	0.000	1.000	-0.075	0.176	
acc_chest_y	-0.124	0.075	1.000	-0.347	
acc_chest_z	0.061	0.178	-0.347	1.000	
electrocardiogram_11	0.007	0.024	0.027	0.010	
electrocardiogram_12	0.011	-0.003	0.007	-0.010	
acc_left_ankle_x	0.077	0.050	-0.028	0.090	
acc_left_ankle_y	-0.007	0.306	-0.092	0.230	
acc_left_ankle_z	0.071	0.102	-0.100	0.235	
gyro_left_ankle_x	0.131	0.043	-0.038	0.048	
gyro_left_ankle_y	0.060	0.248	-0.120	0.250	
gyro_left_ankle_z	0.081	0.050	-0.174	0.230	
MAG_left_ankle_x	0.001	0.092	-0.036	0.107	
MAG_left_ankle_y	0.005	0.011	-0.029	0.023	
MAG_left_ankle_z	-0.001	-0.032	0.035	-0.040	
acc_r_lw_arm_x	-0.002	0.266	0.060	-0.071	
acc_r_lw_arm_y	-0.029	0.193	-0.045	0.309	
acc_r_lw_arm_z	-0.028	0.007	-0.027	0.004	
gyro_r_lw_arm_x	0.087	0.065	0.044	-0.063	
gyro_r_lw_arm_y	-0.001	0.130	-0.149	0.334	
gyro_r_lw_arm_z	0.065	0.073	-0.011	0.003	
MAG_r_lw_arm_x	-0.000	0.010	-0.044	-0.006	
MAG_r_lw_arm_y	0.001	0.012	-0.059	-0.016	
MAG_r_lw_arm_z	-0.002	0.003	0.123	0.029	
label	0.008	0.127	0.094	-0.131	
					electrocardiogram_11 electrocardiogram_12 \
					0.007 0.011
subject					0.024 -0.003
acc_chest_x					0.027 0.007
acc_chest_y					-0.010 -0.010
acc_chest_z					0.010 0.597
electrocardiogram_11					1.000 1.000
electrocardiogram_12					0.597 0.004
acc_left_ankle_x					0.004 -0.002
acc_left_ankle_y					0.003 -0.005
acc_left_ankle_z					0.000 0.000
gyro_left_ankle_x					-0.002 0.013
gyro_left_ankle_y					0.010 0.006
gyro_left_ankle_z					-0.015 -0.012
MAG_left_ankle_x					0.012 0.013
MAG_left_ankle_y					-0.003 0.002
MAG_left_ankle_z					0.008 0.000
acc_r_lw_arm_x					-0.006 -0.011
acc_r_lw_arm_y					-0.016 -0.029
acc_r_lw_arm_z					0.013 0.022
gyro_r_lw_arm_x					0.013 0.022
gyro_r_lw_arm_y					-0.025 -0.033
gyro_r_lw_arm_z					-0.002 -0.003
MAG_r_lw_arm_x					0.010 0.010
MAG_r_lw_arm_y					-0.019 -0.006
MAG_r_lw_arm_z					0.021 0.003
label					0.002 -0.008
					acc_left_ankle_x acc_left_ankle_y acc_left_ankle_z \
					0.077 -0.007 0.071
subject					0.050 0.306
acc_chest_x					-0.028 -0.092
acc_chest_y					0.090 0.230
acc_chest_z					0.004 0.003
electrocardiogram_11					0.002 -0.005
electrocardiogram_12					1.000 0.597
acc_left_ankle_x					0.004 0.003
acc_left_ankle_y					-0.071 0.000
acc_left_ankle_z					0.078 1.000
gyro_left_ankle_x					0.062 0.049
gyro_left_ankle_y					0.087 0.350
gyro_left_ankle_z					-0.034 0.058
MAG_left_ankle_x					-0.031 0.067
MAG_left_ankle_y					-0.223 0.033
MAG_left_ankle_z					0.042 -0.027
acc_r_lw_arm_x					-0.051 0.067
acc_r_lw_arm_y					0.060 0.106
acc_r_lw_arm_z					-0.004 0.092
gyro_r_lw_arm_x					-0.026 0.050
gyro_r_lw_arm_y					0.064 0.152
gyro_r_lw_arm_z					0.016 0.105
MAG_r_lw_arm_x					-0.010 -0.049
MAG_r_lw_arm_y					0.001 -0.061
MAG_r_lw_arm_z					0.005 0.005
label					0.053 0.000
					gyro_left_ankle_x ... acc_r_lw_arm_x acc_r_lw_arm_y \
					0.131 ... 0.002 -0.029
subject					0.043 ... 0.266 0.193
acc_chest_x					-0.038 ... 0.060 0.045
acc_chest_y					0.048 ... -0.071 0.309
acc_chest_z					-0.002 ... -0.006 -0.016
electrocardiogram_11					0.013 ... -0.011 -0.029
electrocardiogram_12					0.062 ... -0.051 0.068
acc_left_ankle_x					0.049 ... 0.067 0.106
acc_left_ankle_y					-0.027 ... 0.012 -0.022
acc_left_ankle_z					0.044 ... -0.061 0.027
acc_r_lw_arm_x					0.015 ... 1.000 -0.166
acc_r_lw_arm_y					-0.004 ... -0.166 1.000
acc_r_lw_arm_z					0.059 ... 0.090 0.145
gyro_r_lw_arm_x					0.368 ... 0.378 -0.189
gyro_r_lw_arm_y					0.027 ... -0.222 0.592
gyro_r_lw_arm_z					0.591 ... 0.100 0.085
MAG_r_lw_arm_x					0.020 ... -0.031 0.030
MAG_r_lw_arm_y					0.027 ... -0.056 0.049
MAG_r_lw_arm_z					-0.030 ... -0.057 0.043
label					0.114 ... -0.001 0.022

Figura 11: Salida parcial del comando de correlación.

Se puede ver que algunos atributos con una correlación de Pearson considerable son:

1. gyro_r_lw_arm_y con acc_r_lw_arm_y: 0.5921
2. electrocardiogram_l1 con electrocardiogram_l2: 0.5965
3. gyro_r_lw_arm_z con gyro_left_ankle_x: 0.591

Según [2] estas medidas entran en un rango de correlación moderado en el ámbito médico.

2.9. Histogramas

Antes de comenzar con el análisis gráfico del dataset, a continuación se deja la liga con todas las imágenes que se obtuvieron en esta parte: Carpeta de imágenes

Para esta parte, se generaron los histogramas de cada atributo del dataset, a excepción de los atributos de label y subject, puesto que estas columnas no relevantes para su análisis con estos gráficos.

```
attributes = data.columns.difference(['subject', 'label'])

# Histogramas generales para cada atributo
for col in attributes:
    color = random_color()
    plt.figure(figsize=(8, 5))
    sns.histplot(data[col], bins='sturges', color=color, edgecolor='black')
    plt.title(f'Histograma de {col}', fontsize=16, fontweight='bold')
    plt.xlabel(col)
    plt.ylabel('Frecuencia')
    plt.savefig(f'{output_dir}/Histograma de {col}.png', bbox_inches='tight', facecolor='white')
    plt.close()
```

A continuación se muestran algunos ejemplos de los gráficos obtenidos:

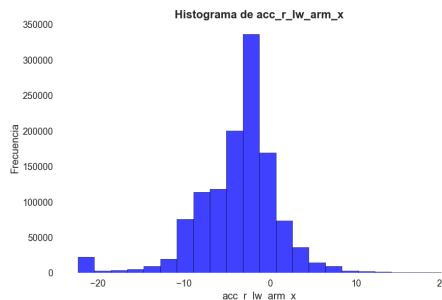


Figura 12: Histograma de la aceleración del sensor del brazo inferior derecho (Eje X)

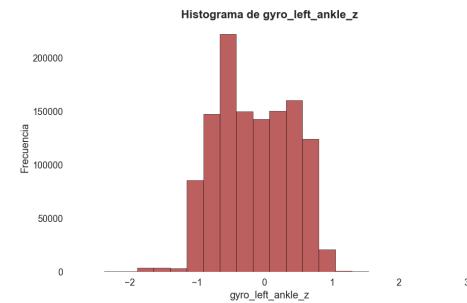


Figura 13: Histograma del giroscopio del sensor del tobillo izquierdo (eje Z)

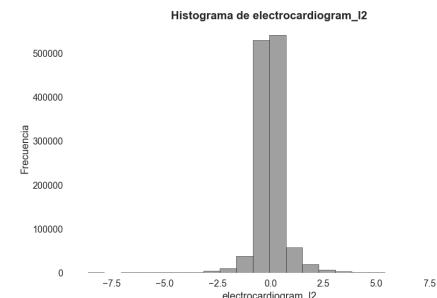


Figura 14: Histograma de la segunda señal de electrocardiograma

Las distribuciones de manera general mostraron que muchas variables tienen distribuciones concentradas alrededor de ciertos valores, lo que indica posibles patrones consistentes entre los sujetos.

Sin embargo, también se observaron colas largas en algunas variables (como en la aceleración o giroscopio), lo que sugiere la presencia de valores atípicos, que pueden significar varias cosas: algún error en los sensores o pueden demostrar las diferentes intensidades a la hora de realizar las actividades físicas (como es en el caso de los electrocardiogramas, donde un valor atípico puede significar algún esfuerzo mayor que a la hora de realizar un ejercicio).

2.10. Gráficas de densidad

El siguiente análisis realizado consistió en la generación de gráficas de densidad para cada atributo del dataset MHEALTH, de nuevo excluyendo las columnas subject y label por las razones ya mencionadas.

```

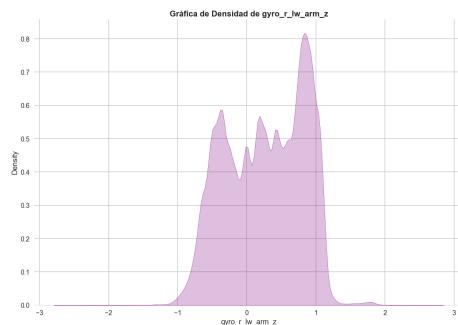
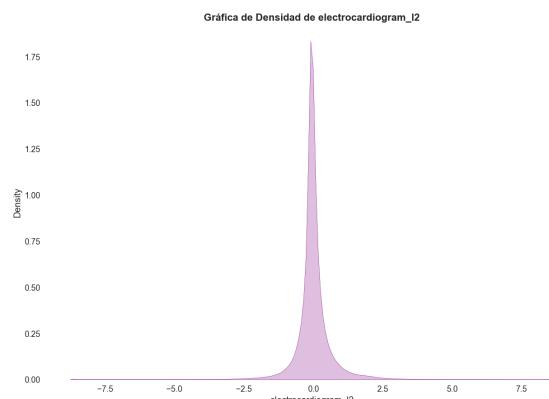
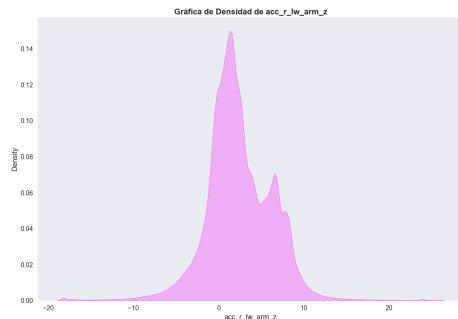
for col in attributes:
    sns.set_style(random_style())
    color = random_color()

    plt.figure(figsize=(12, 8))
    sns.kdeplot(data[col], color=color, fill=True)

    plt.title(f'Gráfica de Densidad de {col}', fontsize=16, fontweight='bold')
    plt.savefig(f'{output_dir}/Densidad de {col}.png', bbox_inches='tight', facecolor='white')
    plt.close()

```

A continuación se muestran algunas de las gráficas obtenidas:



Por ejemplo, los datos de aceleración parecen tener dos o más picos, indicando que los sujetos podrían haber alternado entre movimientos de alta y baja intensidad.

Otras variables, como las relacionadas con el ritmo cardíaco, mostraron colas largas, lo que señala posibles outliers o una variabilidad significativa en las mediciones que son producto, como ya se mencionó, por grandes esfuerzos al momento de realizar los ejercicios.

2.11. Gráfico de caja y bigotes

El siguiente análisis exploratorio consistió en la creación de boxplots (gráficas de cajas y bigotes) para cada atributo, omitiendo, nuevamente, las columnas subject y label.

```

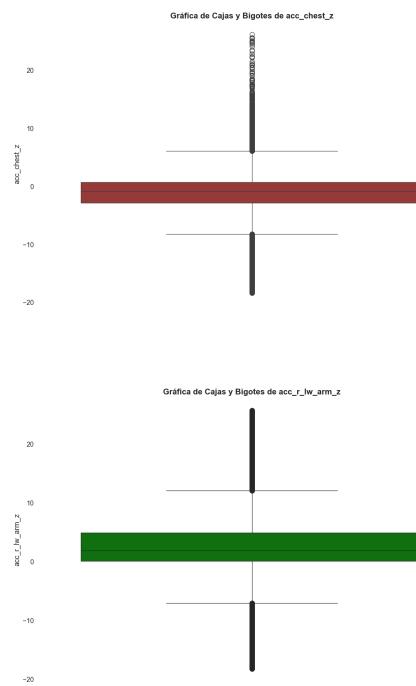
for col in attributes:
    sns.set_style(random_style())
    color = random_color()

    plt.figure(figsize=(12, 8))
    sns.boxplot(y=data[col], color=color)

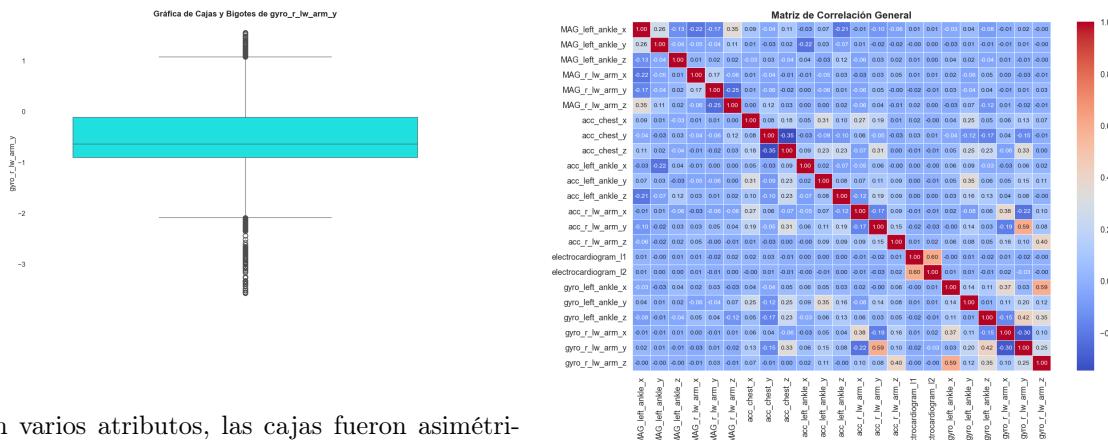
    plt.title(f'Gráfica de Cajas y Bigotes de {col}', fontsize=16, fontweight='bold')
    plt.savefig(f'{output_dir}/Boxplot de {col}.png', bbox_inches='tight', facecolor='white')
    plt.close()

```

A continuación se muestran algunos de los gráficos obtenidos:



Las gráficas revelaron que algunas variables presentan distribuciones unimodales bien definidas, mientras que otras mostraron multimodalidad (varios picos), lo que sugiere la presencia de diferentes estados o actividades físicas dentro de un mismo atributo.



En varios atributos, las cajas fueron asimétricas, lo que indica una distribución sesgada, con valores que tienden a concentrarse más hacia un extremo.

Además, outliers significativos aparecieron en múltiples variables, especialmente en las relacionadas con aceleración y ritmo cardíaco, lo que sugiere que ciertos sujetos realizaron movimientos abruptos o extremos durante las pruebas.

Algunas columnas presentaron rangos muy amplios (cajas grandes), lo que refleja una alta variabilidad, mientras que otras fueron más compactas, señalando atributos más estables.

2.12. Matriz de dispersión

El siguiente paso en el análisis exploratorio consistió en calcular y visualizar la matriz de correlación entre los atributos del dataset, de nuevo, excluyendo los atributos ya mencionados.

De la matriz anterior podemos resaltar lo siguiente:

- Correlaciones positivas más fuertes:

- **electrocardiogram_l1 vs electrocardiogram_l2: 0.60** Existe una correlación positiva considerable entre las dos señales del electrocardiograma, lo que indica que ambas siguen un patrón similar, reflejando posiblemente las respuestas eléctricas sincronizadas del corazón.
 - **gyro_r_lw_arm_y vs gyro_r_lw_arm_z: 0.42** Hay una relación justa entre las rotaciones del brazo derecho en los ejes *y* y *z*. Esto sugiere que ciertos movimientos complejos del brazo afectan ambos ejes de forma coordinada.
 - **Correlación moderada de 0.59:** existe una correlación de 0.59 entre acceleration chest *y* y acceleration chest *z*. Esto sugiere que hay un grado considerable de relación entre los movimientos en esos dos ejes, lo cual tiene sentido biomecánicamente: durante muchos ejercicios — como caminar, correr o incluso sentadillas — el movimiento vertical del pecho (eje *z*) suele estar acompañado por un cambio en el eje *y* debido a la oscilación natural del torso.

Otra correlación cercana a 0.59 ocurre entre gyro left ankle y y gyro left ankle z, lo cual indica que cuando el tobillo rota sobre el eje y (movimientos laterales), también suele producirse un ligero cambio en el eje z (movimientos arriba/abajo). Esto puede deberse a que ciertos ejercicios implican movimientos combinados, como giros de

A continuación la matriz obtenida:

8

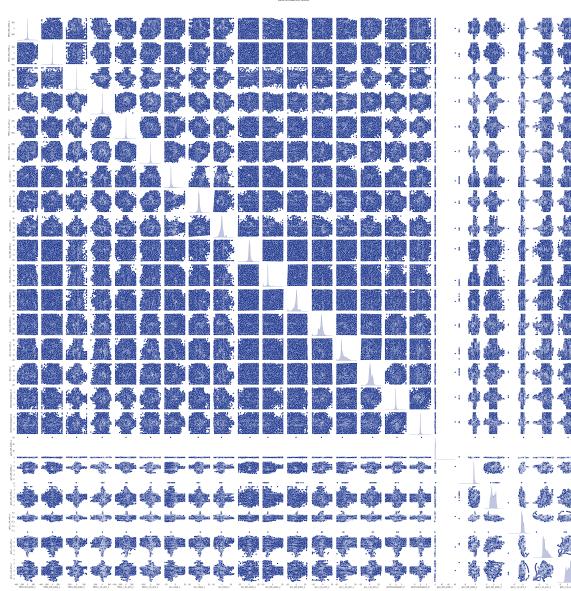
tobillo o cambios de dirección rápidos.

- **acc_left_ankle_y vs acc_left_ankle_z: 0.35** Los movimientos del tobillo izquierdo en los ejes *y* y *z* tienden a estar correlacionados positivamente, probablemente porque los desplazamientos del pie involucran combinaciones de estos dos ejes.

■ Correlaciones negativas más fuertes:

- **acc_left_ankle_x vs acc_r_lw_arm_x: -0.36** La relación negativa entre estos dos valores sugiere que ciertos movimientos en el tobillo izquierdo y el brazo derecho tienden a ocurrir en direcciones opuestas, lo que podría reflejar un patrón biomecánico donde las extremidades superiores e inferiores actúan de manera compensatoria.
- **gyro_left_ankle_y vs gyro_left_ankle_z: -0.25** Hay una correlación negativa moderada entre estos dos ejes del giroscopio en el tobillo izquierdo, lo que indica que ciertas rotaciones tienden a tener componentes inversos en ambos ejes.

Cabe mencionar que estas relaciones no son lo suficientemente fuertes, incluso la literatura las clasifica como pobres o justas, esto según [2].



La mayoría de las gráficas muestran nubes de puntos bastante dispersas, lo que sugiere baja correlación lineal entre muchas variables.

No obstante, hay algunas combinaciones donde los puntos parecen formar patrones más definidos, como ciertas elipses o inclinaciones, lo que podría indicar correlaciones positivas o negativas moderadas.

Por otro lado, algunas relaciones parecen tener formas más estrechas o curvas, lo que indica posibles relaciones no lineales que podrían no capturarse completamente con un análisis solo de correlación de Pearson, y esto además es consistente con los elementos de mayor valor absoluto en la matriz de dispersión.

Las gráficas de dispersión asociadas a giroscopio del tobillo izquierdo en el eje x destacan porque aparecen líneas verticales u horizontales, lo que significa que su valor se mantiene constante independientemente de los valores de las otras variables. Esto sugiere que los valores de este atributo podría haberse registrado sin cambios, indicando un sensor estático o un fallo en la captura de ese eje.

2.13. Matriz de dispersión

Como última parte del análisis gráfico, se generó la matriz de dispersión de los atributos, excluyendo nuevamente los ya mencionados:

```
plt.figure(figsize=(12, 8))
g = sns.pairplot(data[attributes], diag_kind='kde')
g.fig.suptitle('Matriz de Dispersion General', fontsize=16, fontweight='bold', y=1.02)
plt.savefig(f'{output_dir}/Matriz de Dispersion General.png', bbox_inches='tight', facecolor='white')
plt.close()
```

A continuación se muestra la matriz obtenida:

2.14. Escalamiento, normalización o estandarización

La estandarización de las variables del dataset es un paso crucial, especialmente debido a la naturaleza diversa de las mediciones que contiene. Este dataset incluye datos capturados por sensores inerciales (como acelerómetros y giroscopios) así como señales fisiológicas como las provenientes de electrocardiogramas. Dado que estas variables están en diferentes escalas, por ejemplo, las lecturas del electrocardiograma están medidas en milivoltios, mientras que las del acelerómetro están en metros por segundo al cuadrado, no es posible compararlas directamente sin un proceso de estandarización.

La razón por la cual se eligió la estandarización sobre otras técnicas de escalado, como la normalización Min-Max, radica en la naturaleza de los datos. La normalización Min-Max reescalas las variables para que sus valores estén dentro del rango $[0, 1]$, y, como se menciona en [7], puede ser útil si los datos siguen una distribución uniforme o si hay límites claros para cada variable. Sin embargo, este método es muy sensible a los valores atípicos, pues un solo valor extremo puede comprimir drásticamente el rango de los demás datos. Dado que las mediciones del dataset MHEALTH pueden contener ruido o valores inesperados, por ejemplo, debido a movimientos bruscos durante ciertos ejercicios, la normalización Min-Max podría distorsionar la escala de las variables.

En cambio, la estandarización, al convertir las variables a una distribución con media cero y desviación estándar uno, permite que cada característica contribuya de forma equitativa al análisis estadístico y a los algoritmos de aprendizaje automático. Además, es ideal cuando los datos tienen distribuciones gaussianas o desconocidas, como es el caso de las señales fisiológicas y los registros inerciales de este conjunto de datos, como se menciona en [7].

```
# Seleccionar solo las columnas numéricas excluyendo 'subject' y 'label'
numerical_cols = [col for col in data.columns if col not in ['subject', 'label']]

# Crear una copia del dataframe para la estandarización
data_standard = data.copy()

# Aplicar estandarización (Z-score)
standard_scaler = StandardScaler()
data_standard[numerical_cols] = standard_scaler.fit_transform(data[numerical_cols])

# Ver los primeros ejemplos estandarizados
print(data_standard.head())
```

A continuación se muestran las primeras filas del nuevo dataframe obtenido:

	subject	acc_chest_x	acc_chest_y	acc_chest_z	electrocardiogram_11	\
0	1	-0.3180	0.1047	0.3782	0.0125	
1	1	-0.3255	0.3450	0.3999	0.0125	
2	1	-0.2792	0.1851	0.4178	0.0293	
3	1	-0.2769	0.2002	0.3627	0.1134	
4	1	-0.2897	0.2421	0.3826	0.3040	
	electrocardiogram_12	acc_left_ankle_x	acc_left_ankle_y	acc_left_ankle_z		
0	0.0120	0.1805	-0.0009	0.2903		
1	0.0292	0.2335	0.0443	0.3000		
2	0.0580	0.2390	0.0301	0.2995		
3	0.1674	0.1796	0.0630	0.2756		
4	0.2883	0.2412	0.0729	0.3050		
	gyro_left_ankle_x	...	acc_r_lw_arm_x	acc_r_lw_arm_y	acc_r_lw_arm_z	\
0	0.2148	...	-1.0363	0.2132	-0.5691	
1	0.1770	...	-1.0316	0.2580	-0.6114	
2	0.1770	...	-1.0060	0.2654	-0.5464	
3	0.1770	...	-1.0317	0.2587	-0.5227	
4	0.1770	...	-1.0470	0.2882	-0.5125	
	gyro_r_lw_arm_x	gyro_r_lw_arm_y	gyro_r_lw_arm_z	MAG_r_lw_arm_x	\	
0	-0.3277	-0.9790	-0.4113	-0.0945		
1	-0.3277	-0.9790	-0.4113	-0.0875		
2	-0.3277	-0.9790	-0.4113	-0.0670		
3	-0.3425	-0.9752	-0.4266	-0.0468		
4	-0.3425	-0.9752	-0.4266	-0.0265		
	MAG_r_lw_arm_y	MAG_r_lw_arm_z	label			
0	-0.0695	0.0048	0			
1	-0.0478	0.0100	0			
2	-0.0263	0.0048	0			
3	-0.0212	-0.0002	0			
4	-0.0161	-0.0105	0			

[5 rows x 25 columns]

Cabe destacar, que en 2 de los 3 modelos seleccionados que se presentarán más adelante, es requerida la estandarización de los datos para su correcto funcionamiento.

2.15. Características Extraídas

Fueron generadas 164 características provenientes de 3 de los sensores utilizados: Acelerómetro, electrocardiograma y giroscopio.

2.15.1. Acelerómetro

Para la extracción de características de los datos provenientes de los acelerómetros, tomaremos como guía lo mencionado en [8], tanto en el tamaño de las ventanas de tiempo (3 segundos sin overlapping), como en las características por extraer de los sensores:

- Media de cada uno de los 3 ejes (X, Y, Z).
- Desviación estándar de cada uno de los 3 ejes.
- Valor máximo de cada uno de los 3 ejes.

- Correlación entre cada par de ejes (XY, XZ, YZ).
- Media de la magnitud de la señal.
- Desviación estándar de la magnitud de la señal.
- Área bajo la curva (AUC) de la magnitud (según la ecuación 1 del documento).
- Diferencias medias de la magnitud entre lecturas consecutivas (según la ecuación 2 del documento).
- Magnitud de la señal, que representa la contribución total de la aceleración de los 3 ejes (según la ecuación 3 del documento).

$$AUC = \sum_{t=1}^T magnitude(t)$$

$$meandif = \frac{1}{T-1} \sum_{t=2}^T (magnitude(t) - magnitude(t-1))^2$$

$$Magnitude(x, y, z, t) = \sqrt{a_x(t)^2 + a_y(t)^2 + a_z(t)^2}$$

2.15.2. ECG

A pesar de que en el proyecto original en el que fue usado el dataset, no fue utilizada la información proveniente del ECG, nosotros consideramos su inclusión debido a que en la literatura se encontró que la fusión de datos provenientes de acelerómetros y ECG, en tareas de reconocimiento de la actividad humana, puede traer consigo mejoras significativas en el desempeño del modelo.

En Human Activity Recognition: A Comparative Study to Assess the Contribution Level of Accelerometer, ECG, and PPG Signals [9], se encontró específicamente que, el uso de ECG y Acelerómetro, en comparación con el uso de solo el acelerómetro, ayuda al modelo a clasificar de mejor manera entre actividades estacionarias y no estacionarias. Recalcando para nuestro caso, la mejora vista en la clasificación de las actividades 'caminar' y 'subir escaleras', actividades presentes en nuestro dataset.

Debido a lo anterior, también se tomó la decisión de replicar parcialmente los features utilizados en la literatura, para tratar de replicar los resultados obtenidos.

Características Extraídas:

- **Media:** valor promedio de los datos.
- **Mínimo:** el valor más pequeño.
- **Máximo:** el valor más grande.

- **Mediana:** el valor correspondiente al percentil 50 %.
- **Desviación estándar:** mide qué tan dispersos están los datos respecto al valor promedio.
- **Tasa de cruce por cero:** cuenta cuántas veces la serie temporal cruza la línea $y = 0$.
- **Tasa de cruce por la media:** cuenta cuántas veces la serie temporal cruza la línea $y = \mu$.

2.15.3. Giroscopio

De los datos obtenidos de los giroscopios, utilizaremos las características mencionadas en [10].

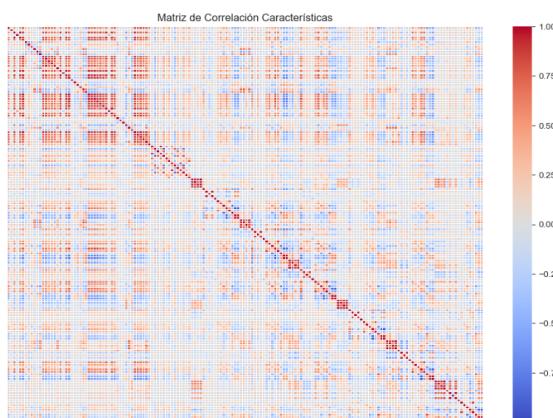
Esto se debe a que, en dicho artículo, se encontró que al utilizar el giroscopio junto con el acelerómetro se pueden obtener mejores resultados de clasificación para algunas actividades que también están presentes en nuestro conjunto de datos: Sentado, De pie y Subiendo escaleras.

Características Extraídas:

- **Dominio del tiempo:** Interquartile range, Max, Mean, Median, Min, Root mean square, Skewness, Standard deviation, Absolute energy, Autocorrelation, Centroid Entropy, Zero crossing rate, Histogram
- **Dominio de la frecuencia:** Fundamental frequency, Max power spectrum, Maximum frequency, Median frequency, Spectral entropy, Spectral kurtosis, Spectral skewness

2.15.4. Matriz de correlación

La matriz de correlación de todas las características obtenidas, resultó de la siguiente manera:



Es fácil notar que hay valores de correlación absolutos muy grandes, entre ellos valores de hasta 0.97 de correlación, lo que indica que contamos

con muchas características que son redundantes. Abriendo la posibilidad para un proceso de reducción de características.

2.16. Modelos de Aprendizaje

Fueron implementados los modelos Random Forest, Logistic Regression y Perceptrón, esto basado en sus pros y contras en el área del reconocimiento de la actividad humana encontrados en diversas fuentes.

2.16.1. Random Forest

Los bosques aleatorios para regresión y clasificación se encuentran actualmente entre los métodos de aprendizaje automático más utilizados. Son muy potentes, a menudo funcionan bien sin un ajuste excesivo de los parámetros y trabajan bien con datos con variabilidad en los rangos, de modo que no es necesario reescalar. Al igual que evitan en cierta medida el sobreajuste que suelen provocar los árboles de decisión por sí solos. Müller, A. & Guido S., (2017) [11].

Asimismo, en Reiman, L. (2001) [12]. Se menciona que los random forest son robustos al ruido en los datos, y como se ha visto en secciones anteriores, en los datos provenientes de sensores como los utilizados en la generación de este dataset es muy común encontrarnos con esa situación.

De igual forma, en el artículo de Zaki, Z., Shah, M. A., Wakil, K.,& Sher, F. (2020) [13], en el que el Random Forest se puso a prueba en conjunto con otros modelos de clasificación sobre los datasets UCI-HAR y HAPT, este obtuvo más del 90 % de accuracy en ambos.

2.16.2. Logistic Regression

La regresión logística, y en general los modelos lineales son muy rápidos para entrenar y predecir, al igual que suelen tener un buen performance en situaciones en el que el número de características es muy grande comparado con el número de muestras [11], caso similar al de este dataset, en el que tenemos más de 100 características y los ejemplos por clase son en promedio 200.

Asimismo, otro factor a tomar en cuenta para seleccionar el uso de este modelo fue lo encontrado en [13], en el que la regresión logística se puso a prueba en conjunto con otros modelos de clasificación sobre los datasets UCI-HAR y HAPT, en

los que obtuvo una accuracy del 96.1 % y 94.5 % respectivamente.

2.16.3. Perceptrón

El Perceptrón es un algoritmo de aprendizaje automático lineal simple que puede ser efectivo para problemas de clasificación multiclas. Aunque es menos sofisticado que otros modelos, puede servir como línea base.

El Perceptrón, aunque simple, puede ser sorprendentemente efectivo para clasificación de actividades humanas cuando las características están bien seleccionadas. Según Bulling et al. (2014) [14] en su trabajo sobre reconocimiento de actividad humana, los modelos lineales como el Perceptrón pueden alcanzar buenos resultados en conjuntos de datos con patrones claramente separables, como movimientos corporales distintos capturados por sensores.

2.16.4. Desempeño de modelos base

Para medir el rendimiento de los modelos, se utilizó la técnica de validación cruzada.

Recordemos que la validación cruzada es un método estadístico de evaluación del rendimiento de la generalización más estable y exhaustivo que la división en un conjunto de entrenamiento y otro de prueba. [11]

Especificamente, utilizó la variación de validación cruzada estratificada con grupos (Stratified-GroupKFold), en la que se tomarán folds con muestras de un solo grupo y cada fold será usado tanto como para entrenamiento como para prueba, pero sin aparecer en ambos durante la misma iteración; recordando que este dataset contiene información de 10 personas distintas (grupos).

Asimismo, utilizaron las métricas Accuracy y F1-macro:

- Accuracy: La fracción de ejemplos clasificados correctamente.
- F1-macro: La métrica más utilizada para los conjuntos de datos desequilibrados en la configuración multiclas, es la versión multiclas del F1-Score. Calcula los F-scores no ponderados por clase. Esto da el mismo peso a todas las clases, sin importar cuál sea su tamaño.

De esta forma, los resultados obtenidos fueron los siguientes:

```

    === Métricas de TEST (validación) ===
Accuracy por fold: [0.9 0.9498 0.9738 0.9124 0.8869 0.973 0.9048 0.9955 0.9214 0.9183]
F1-macro por fold: [0.8764 0.9475 0.9752 0.8947 0.8673 0.9744 0.9026 0.9958 0.9094 0.921 ]
Accuracy promedio (test): 0.9336
F1-macro promedio (test): 0.9264

    === Métricas de TRAIN ===
Accuracy por fold (train): [0.9905 0.9796 0.9815 0.9796 0.9761 0.9711 0.981 0.9766 0.9815 0.9807]
F1-macro por fold (train): [0.9912 0.981 0.9828 0.981 0.9776 0.9725 0.9822 0.9781 0.9827 0.9819]

Accuracy promedio (train): 0.9798
F1-macro promedio (train): 0.9811

```

Figura 15: Resultados del modelo Random Forest en el dataset MHEALTH.

En general, podemos ver que nuestro modelo generaliza bien, pues en cada fold se dan F1 score y Accuracy superiores a 0.85, manteniendo un promedio de 0.93 y 0.92 respectivamente.

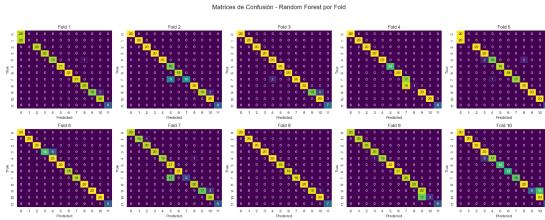


Figura 16: Matrices de confusión del Random Forest generadas por cada fold.

Asimismo, verificando las matrices de confusión generadas en cada iteración, podemos darnos cuenta de que para ciertos folds el modelo no logra identificar bien la separación entre las actividades que son muy similares entre sí, como:

L1: Mantenerse de pie / L2: Sentado y relajado al igual que:

L10: Trotando / L11: Corriendo

Las cuales son pares de actividades que comparten muchas similitudes en aspectos como forma del movimiento, nivel de fuerza de la actividad, entre otros.

Por otra parte, la regresión logística nos dio los siguientes resultados:

```

    === Métricas de TEST (validación) ===
Accuracy por fold: [0.9598 0.9954 0.8646 0.8965 0.8462 0.9595 0.8225 0.9955 0.9301 0.7067]
F1-macro por fold: [0.9708 0.9957 0.8428 0.7698 0.8248 0.9467 0.7738 0.9958 0.9206 0.7125]

Accuracy promedio (test): 0.8897
F1-macro promedio (test): 0.8753

    === Métricas de TRAIN ===
Accuracy por fold (train): [0.951 0.9523 0.9575 0.9622 0.9403 0.9532 0.9505 0.9521 0.9545 0.9575]
F1-macro por fold (train): [0.9538 0.9544 0.9595 0.9638 0.9432 0.9559 0.953 0.9549 0.9568 0.9598]

Accuracy promedio (train): 0.9531
F1-macro promedio (train): 0.9555

```

Figura 17: Resultados del modelo Logistic Regression en el dataset MHEALTH.

De igual manera, su rendimiento promedio en

los test: 0.88 en Accuracy y 0.87 en F1 score nos indican que el modelo es capaz de generalizar de manera satisfactoria la mayoría de las veces.

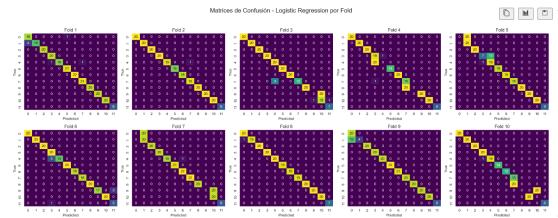


Figura 18: Matrices de confusión de Logistic Regression generadas por cada fold.

Podemos notar que este modelo tiende a fallar en la clasificación de las mismas actividades que el Random Forest, al igual que, en ocasiones no diferencia bien entre las actividades caminar y subir escaleras.

Por último, los resultados del Perceptrón fueron los siguientes:

```

    === Métricas de TEST (validación) ===
Accuracy por fold: [0.987 0.9726 0.8777 0.8894 0.9095 0.9324 0.8139 0.9821 0.8821 0.8173]
F1-macro por fold: [0.9874 0.9741 0.8669 0.8956 0.9051 0.9223 0.7694 0.9832 0.8498 0.7969]

Accuracy promedio (test): 0.9064
F1-macro promedio (test): 0.8945

    === Métricas de TRAIN ===
Accuracy por fold (train): [0.9835 0.9692 0.9935 0.9841 0.9751 0.9796 0.989 0.9641 0.9675 0.9832]
F1-macro por fold (train): [0.9807 0.9705 0.9934 0.984 0.976 0.9803 0.9865 0.9652 0.9687 0.9837]

Accuracy promedio (train): 0.9789
F1-macro promedio (train): 0.9789

```

Figura 19: Resultados del modelo Perceptron en el dataset MHEALTH.

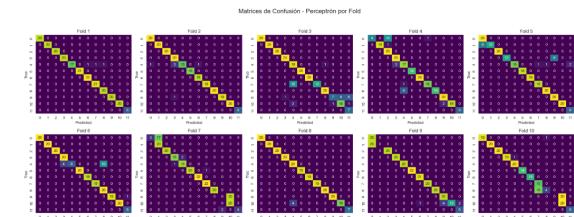


Figura 20: Matrices de confusión del Perceptron generadas por cada fold.

Podemos notar que a este modelo, para algunas iteraciones, también repite los errores de los modelos pasados, pero también presenta dificultades para diferenciar entre la actividad de agacharse y manejar bicicleta en algunos folds, esto puede deberse a que la posición que toman nuestras rodillas y piernas durante el manejo de la bicicleta es

similar a la que se toma para agacharse. De igual forma presenta ligera variabilidad de resultados dependiendo de cada fold.

Sin embargo, es importante señalar que, en general, este modelo tuvo un rendimiento superior a la regresión logística, pues en promedio tuvo 0.90 de Accuracy y 0.89 de F1-macro, lo que indica que es ligeramente superior en cuanto a la capacidad de generalizar para datos no vistos durante el entrenamiento.

2.16.5. Conclusiones del desempeño

Hasta este punto, el Random Forest fue el modelo que generó los mejores resultados (0.93 y 0.92 de Accuracy y F1 en promedio).

De igual forma podemos notar que los 3 modelos tuvieron dificultades para diferenciar entre actividades con una inercia similar. Lo cual estaba dentro de los problemas esperados mencionados en [15].

2.17. Métodos de reducción de características

La reducción de dimensionalidad es crucial para mejorar la eficiencia computacional y potencialmente el rendimiento del modelo (Guyon & Elisseeff, 2003) [28]. Implementamos tres técnicas:

2.17.1. Selección de Características con PCA

PCA es una técnica lineal que proyecta los datos en un espacio de menor dimensionalidad maximizando la varianza (Jolliffe, 2002)[29].

PCA es particularmente útil para datos de sensores de actividad humana. Según Preece et al. (2009) [30], las señales de acelerómetros y giroscopios suelen tener alta correlación entre canales, lo que hace que PCA sea efectivo para reducir dimensionalidad manteniendo la información relevante.

2.17.2. Selección de Características con Kernel PCA

Kernel PCA es una extensión no lineal de PCA que puede capturar relaciones más complejas (Schölkopf et al., 1997). [34]

Justificación según la Literatura (Mika et al., 1999):

-Selección de Componentes: En la práctica, para problemas de reconocimiento de actividad humana, se recomienda seleccionar un número fijo de componentes basado en: La capacidad computacional disponible, la regla empírica de mantener al menos el 95 % de la varianza (aunque para KPCA esto es menos directo que en PCA), parámetro gamma: El valor de gamma en el kernel RBF controla el suavizado. Para datos de sensores de actividad humana, estudios como el de Zhang et al. (2012) sugieren valores entre 0.01 y 1.0.

Para el kernel RBF, calculamos gamma automáticamente adaptado a la escala de nuestros datos, como sugirió Zhang et al. (2016) para datos de wearables:

$$GV = 1/(X_train.shape[1] \times X_train.values.var())$$

Implementamos KPCA con parámetros optimizados basados en la literatura (Mika et al., 1999; Hammerla et al., 2016). [35] [40]

2.17.3. Selección de Características con ANOVA F-test

En aplicaciones de mHealth (ej. clasificación de movimientos, detección de caídas), los datos provienen de múltiples sensores (acelerómetros, giroscopios, etc.), generando un alto número de características. El ANOVA F-test evalúa si las medias de estas características difieren significativamente entre clases (ej. "caminar" vs. "correr"), permitiendo seleccionar las más discriminativas (Kira & Rendell, 1992).

Los datos de actividad humana suelen ser de alta dimensionalidad (ej. ventanas temporales, frecuencias, ejes XYZ). Estudios como el de Anguita et al. (2013) en el dataset Human Activity Recognition (HAR) demostraron que el ANOVA F-test ayuda a eliminar características redundantes, mejorando el rendimiento de modelos como SVM y Random Forest.

2.17.4. Resultados de los métodos

Podemos observar que tanto la técnica de PCA como la de Kernel PCA, únicamente mejoraron ligeramente las métricas para el modelo de Regresión Logística, sin embargo para los otros modelos inclusive redujeron su performance.

Por otro lado, la técnica de Anova-F mejoró significativamente el Accuracy y F1 del Random Fo-

rest y de la Regresión Logística, el Accuracy promedio pasó de 0.93 a 0.95 en el Random Forest, en la Regresión Logística de 0.89 a 0.81 y el F1 promedio pasó de 0.93 a 0.94 en el Random Forest y de 0.88 a 0.90 en la regresión logística, en tanto que para el Perceptrón, empeoró en ambas métricas.

Otro aspecto remarcable del método antes mencionado, es que fue el que redujo a la menor cantidad de características, pasando de 164 a 50, en contraste a 55 y 100 del PCA y Kernel PCA respectivamente. Por lo que podemos notar que el Anova F-test seleccionó las características más representativas del conjunto.

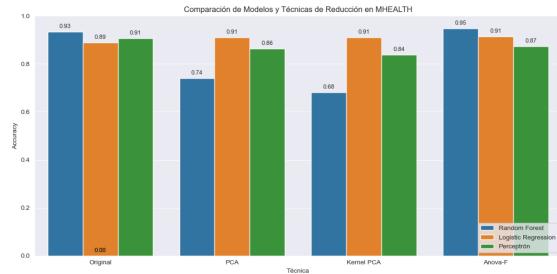


Figura 21: Efecto de los métodos sobre el Accuracy.

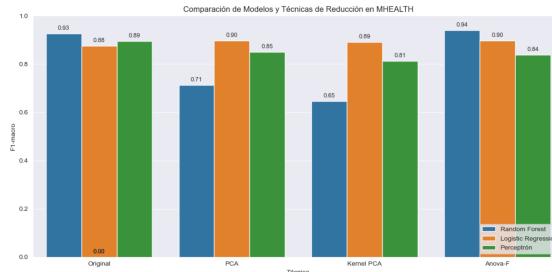


Figura 22: Efecto de los métodos sobre el F1-score.

2.18. Métodos de fusión

Las técnicas de aprendizaje automático mediante ensamblado consisten en combinar múltiples modelos para mejorar el rendimiento predictivo. Estas técnicas han surgido como metodologías poderosas para el modelado predictivo y el análisis de datos.

Las técnicas de ensamblado como voting, bagging y boosting mejoran significativamente el rendimiento de los modelos de aprendizaje automático, superando así a los modelos individuales o

tradicionales. Se destaca la importancia de estas técnicas para abordar errores del modelo como el sesgo y la varianza. Mientras que bagging busca reducir la varianza, boosting se enfoca en minimizar el sesgo. [31]

Notemos que para el caso del bagging, un ejemplo común es el Random Forest, el cual ya ha sido usado como modelo base en secciones anteriores y ha demostrado ser el que mejores resultados ha obtenido. Es por esto que para los ensambles Voting y Boosting, se compararán sus rendimientos con el del Random Forest.

2.18.1. Voting Ensamble

Los métodos de ensamble por votación combinan las predicciones de múltiples clasificadores base, reduciendo el sesgo individual y mejorando la precisión general en HAR. Esto es especialmente útil cuando se trabaja con datos de sensores multi-modal, donde diferentes modelos capturan distintos aspectos de las señales. [14]

En el caso del Weighted Majority Vote (seleccionado para este dataset), se le asignan pesos específicos a cada modelo dentro del ensamble. En este caso se utilizaron los modelos base mencionados anteriormente, al igual que se utilizó la reducción de características con ANOVA F-test. De igual forma, los pesos fueron 0.6 para el Random Forest, y 0.5 para los otros 2 modelos.

2.18.2. Boosting

Los métodos de boosting, como AdaBoost y Gradient Boosting, mejoran significativamente la precisión en el reconocimiento de actividades humanas al combinar múltiples modelos débiles (ej: árboles de decisión poco profundos) en un predictor fuerte, reduciendo tanto el sesgo como la varianza.

Resultados experimentales han demostrado la viabilidad de los clasificadores de ensamble AdaBoost al lograr un mejor rendimiento para el reconocimiento automatizado de actividades humanas mediante el uso de sensores corporales. Los resultados han mostrado que los clasificadores de ensamble basados en el algoritmo Adaboost mejoran significativamente el rendimiento del reconocimiento automatizado de actividades humanas (HAR). [32]

Para este caso se utilizó un Ensamble AdaBoost, con arboles de decisión de profundidad 3

como estimadores base.

2.18.3. Comparación del desempeño

Los resultados obtenidos fueron:

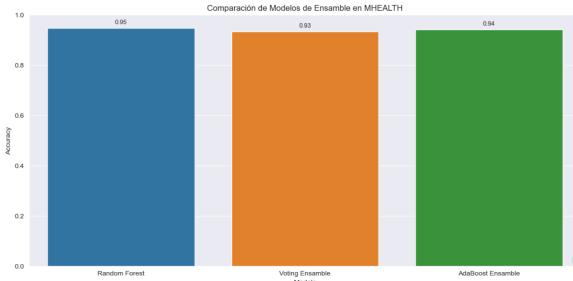


Figura 23: Accuracy de los ensambles

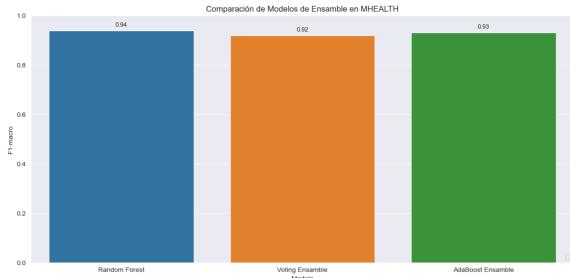


Figura 24: F1-macro de los ensambles

Podemos notar que ni el Voting ni el Boosting lograron superar el rendimiento obtenido por el Random Forest (Bagging Ensamble) con características reducidas.

2.19. Conclusiones Finales

Para el dataset MHEALTH, el mejor modelo base resultó ser el Random Forest, con Accuracy y F1-score promedios de 0.93 y 0.92 respectivamente. Seguido por el perceptrón con 0.9064 y 0.8945 y la Regresión Logística con 0.8897 y 0.8753 respectivamente.

Ahora, en cuanto al mejor método de reducción de características encontramos el Anova F-test, que redujo las características de 164 a 50 y aumentó el accuracy del Random Forest de 0.93 a 0.95 y el F1 de 0.93 a 0.94.

Por último, como se mencionó en la sección anterior, el mejor método de Ensamble para el MHEALTH fue el Bagging, específicamente implementado mediante un Random Forest.

3. UTD Multimodal Human Action Dataset

3.1. Descripción del dataset

El reconocimiento de acciones humanas a partir de sensores ha ganado relevancia en aplicaciones como la interacción humano-computadora. La combinación de sensores multimodales, como cámaras de profundidad y sensores iniciales, permite una caracterización más robusta del movimiento humano. En este contexto, el presente trabajo utiliza el Conjunto de Datos de Acción Humana Multimodal de la Universidad de Texas en Dallas (UTD-MHAD) [1], que integra datos de una cámara Kinect y un sensor inercial portátil. Con respecto a los sensores, la cámara Kinect puede capturar imágenes a color con una resolución de 640 x 480 pixeles y una imagen de profundidad de 16 bits con una resolución de 320 x 240. La tasa de frames de este dispositivo es de 30 frames por segundo [2]. Igualmente, el sensor inercial portátil utilizado para esta base de datos consta de un sensor MEMS de 9 ejes que captura aceleración de 3 ejes, velocidad angular de 3 ejes y fuerza magnética de 3 ejes. La tasa de muestreo de este sensor es de 50 Hz [3].

La base de datos consiste en 27 diferentes acciones:

1. Deslizamiento del brazo derecho hacia la izquierda
2. Deslizamiento del brazo derecho hacia la derecha
3. Saludo con la mano derecha
4. Aplauso frontal con ambas manos
5. Lanzamiento con el brazo derecho
6. Cruzar los brazos sobre el pecho
7. Tiro de baloncesto
8. Dibujar una "X" con la mano derecha
9. Dibujar un círculo con la mano derecha (en sentido horario)
10. Dibujar un círculo con la mano derecha (en sentido antihorario)
11. Dibujar un triángulo
12. Lanzamiento de bolos (mano derecha)
13. Boxeo frontal
14. Swing de béisbol desde la derecha
15. Golpe de derecha en tenis con la mano derecha
16. Curl de brazos (ambos brazos)
17. Saque de tenis
18. Empuje con ambas manos

19. Golpear una puerta con la mano derecha
20. Atrapar un objeto con la mano derecha
21. Recoger y lanzar con la mano derecha
22. Trotar en el lugar
23. Caminar en el lugar
24. Sentarse desde posición de pie
25. Levantarse desde posición sentada
26. Estocada hacia adelante (pie izquierdo adelante)
27. Sentadilla (con ambos brazos extendidos hacia adelante)

El sensor inercial se colocó en la muñeca derecha o en el muslo derecho del sujeto, dependiendo de si la acción era principalmente de brazo o pierna. Específicamente, para las acciones del 1 al 21, el sensor inercial se colocó en la muñeca derecha y para las acciones del 22 al 27, el sensor inercial se colocó en el muslo derecho del sujeto. La base de datos contiene 27 acciones realizadas por 8 personas (4 hombres y 4 mujeres). Cada persona repite las 27 acciones. Despues de remover 3 secuencias corruptas, la base de datos incluye 861 secuencias de datos. Se grabaron 4 tipos de datos: Videos en RGB, Videos de profundidad, posiciones de las articulaciones del esqueleto y las señales del sensor inercial en tres ejes.

1. Cabeza
2. Centro de los hombros
3. Espina dorsal
4. Centro de la cadera
5. Hombro izquierdo
6. Codo izquierdo
7. Muñeca izquierda
8. Mano izquierda
9. Hombro derecho
10. Codo derecho
11. Muñeca derecha
12. Mano derecha
13. Cadera izquierda
14. Rodilla izquierda
15. Tobillo izquierdo
16. Pie izquierdo
17. Cadera derecha
18. Rodilla derecha
19. Tobillo derecho
20. Pie derecho

Cada dato de esqueleto es una matriz en donde cada fila de un fotograma del esqueleto corresponde a tres coordenadas espaciales de una articulación.

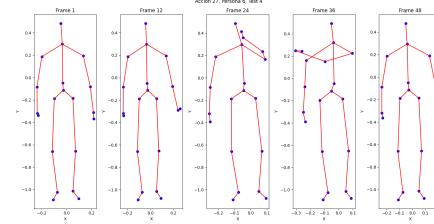


Figura 26: esqueleto visualización

			point_1_x	point_1_y	point_1_z	\	
action_id	person_id	test_id	frame_idx	-0.045763	0.483985	2.830189	
1	1	1	0	-0.045314	0.483987	2.830307	
			1	-0.044791	0.483999	2.830396	
			2	-0.044316	0.483995	2.830455	
			3	-0.043912	0.483994	2.830503	
			4				
			point_2_x	point_2_y	point_2_z	\	
action_id	person_id	test_id	frame_idx	-0.034567	0.299454	2.877848	
1	1	1	0	-0.034420	0.299377	2.877829	
			1	-0.034372	0.299287	2.876948	
			2	-0.034331	0.299194	2.876836	
			3	-0.034305	0.299102	2.876704	
			4				
			point_3_x	point_3_y	point_3_z	\	
action_id	person_id	test_id	frame_idx	-0.030579	-0.049345	2.890448	
1	1	1	0	-0.030435	-0.049361	2.890594	
			1	-0.030323	-0.049369	2.890700	
			2	-0.030229	-0.049374	2.890781	
			3	-0.030150	-0.049374	2.890841	
			4				

Figura 25: Datos de puntos esqueleto

La última modalidad corresponde a las señales del sensor inercial, que incluyen las señales de aceleración en tres ejes y de rotación en tres ejes.

			acc_x	acc_y	acc_z	\	
action_id	person_id	test_id	frame_idx	-0.959473	-0.177734	-0.192871	
1	1	1	0	-0.961914	-0.153320	-0.159912	
			1	-0.974609	-0.152832	-0.145996	
			2	-0.941895	-0.135742	-0.127938	
			3	-0.958252	-0.201416	-0.139484	
			4				
			gyro_x	gyro_y	gyro_z	\	
action_id	person_id	test_id	frame_idx	5.221374	1.536718	0.153672	
1	1	1	0	6.778626	1.954198	0.244275	
			1	11.267176	3.175573	1.099237	
			2	16.885496	4.732824	2.320611	
			3	16.030534	4.000000	0.366412	
			4				

Figura 27: Datos iniciales

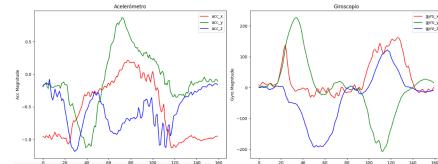


Figura 28: Visualizacion de datos iniciales

3.2. Identifica y elimina datos ausentes

En este caso, el creador de la base de datos menciona que hubo tres archivos corruptos; sin embargo, estos fueron eliminados previamente. Según el

autor, todos los demás archivos se encuentran en perfectas condiciones.

3.3. Tipos de datos de cada atributo

point_2_x	float64
point_2_y	float64
point_2_z	float64
point_3_x	float64
point_3_y	float64
point_3_z	float64

Figura 29: tipo de dato esqueleto

acc_x	float64
acc_y	float64
acc_z	float64
gyro_x	float64
gyro_y	float64
gyro_z	float64
dtype:	object

Figura 30: tipo de dato inercial

Este `dtype` indica que todas las columnas contienen números en punto flotante (`float64`), lo cual es ideal para el análisis y modelado de datos iniciales y de esqueleto

3.4. Resumen estadístico

A continuación, se obtiene el resumen estadístico de cada uno de los atributos de UTD-MHAD Tabla 2. Estadísticas descriptivas de los datos iniciales (acelerómetro y giroscopio)

Métrica	acc_x	acc_y	acc_z	gyro_x	gyro_y	gyro_z
count	155638	155638	155638	155638	155638	155638
mean	-0.6587	-0.2832	-0.0351	2.8892	-4.7958	-0.1656
std	0.7513	0.5308	0.5637	97.8789	112.0750	122.6850
min	-8.0000	-8.0000	-8.0000	-1000.5496	-1000.5496	-741.5573
25%	-1.0103	-0.5825	-0.3427	-26.8702	-25.3740	-33.7099
50%	-0.9319	-0.2249	0.0305	-0.8855	0.7023	-0.8855
75%	-0.2244	0.0005	0.2896	36.7023	20.6412	37.8550
max	3.6528	7.7253	6.3982	1000.5191	606.7786	1000.5191

Con respecto a la tabla 1, los datos iniciales provenientes del acelerómetro y giroscopio del conjunto de datos UTD-MHAD revela características fundamentales del comportamiento dinámico asociado a las distintas acciones humanas. En particular, se observa que la media del acelerómetro en los tres ejes (acc_x, acc_y, acc_z) se encuentran cercanos a cero, con un leve sesgo negativo en los ejes X e Y. Esto es consistente con estudios previos [4] que señalan que la aceleración media durante tareas humanas tiende a aproximarse a cero cuando se considera un conjunto diverso de acciones, ya que los movimientos positivos y negativos tienden a compensarse a lo largo del tiempo. Las desviación estándar en el acelerómetro se encuentran entre 0.5 y 0.75, lo que indica una variabilidad moderada en la magnitud de los movimientos. Esta variabilidad es esencial para diferenciar acciones que requieren más esfuerzo físico o mayor movilidad, como "jogging." "tennis serve", de otras más estáticas como "cross arms." "stand to sit". En contraste, los valores del giroscopio muestran una dispersión mucho mayor, con desviaciones estándar superiores a 97 en todos los ejes. Esta alta variabilidad angular refleja la riqueza cinemática del conjunto de datos y sugiere que los giroscopios capturan con mayor sensibilidad las diferencias en las velocidades rotacionales de las acciones.

Tabla 3. Estadísticas descriptivas de los puntos esqueletos

Punto	Media	Desviación estándar	Mínimo	Máximo
point_1_x	-0.095496	0.052126	-0.344944	0.192717
point_1_y	0.453867	0.163035	-0.584201	0.822258
point_1_z	2.881422	0.127305	2.250838	3.584473
point_2_x	-0.090920	0.045769	-0.304446	0.153682
point_2_y	0.268885	0.156211	-0.696218	0.654998
point_2_z	2.899522	0.124685	2.329529	3.573770
point_3_x	-0.084452	0.039323	-0.273464	0.135600
point_3_y	-0.070403	0.128158	-0.857664	0.281258
point_3_z	2.916187	0.117721	2.390860	3.500698

El análisis estadístico permite inferir patrones generales en la postura y desplazamiento del cuerpo humano durante las actividades registradas. A partir de los valores se pueden extraer conclusiones relevantes sobre la disposición espacial del cuerpo. En cuanto al eje Z presenta en general valores de media entre 2.7 y 3.0 unidades para la mayoría de los puntos, lo cual sugiere que los sujetos se encontraban a una distancia relativamente constante del sensor o cámara. En contraste, los ejes X y Y muestran mayor variabilidad entre puntos, reflejando la disposición anatómica del cuerpo y la cinemática de cada extremidad. En cuanto a la justificación de los atributos, los atributos seleccionados del conjunto UTD-MHAD [1] se basan en información esquelética 3D y datos iniciales por su capacidad para representar de forma complementaria la postura y dinámica del cuerpo humano. Las coordenadas (x, y, z) de las articulaciones capturan la estructura espacial de cada acción, mientras que las velocidades, aceleraciones y ángulos articulares reflejan la variación temporal y las relaciones internas del movimiento. Estos atributos fueron elegidos por su efectividad para discriminar acciones con base en forma, ritmo e intensidad. A continuación, se examina la relación entre las coordenadas articulares tridimensionales y los datos provenientes de sensores iniciales (acelerómetro y giróscopo) con el objetivo de identificar patrones de co-movimiento.

3.5. Distribución de las clases

action_id	person_id	test_id	
1	1	1	160
		2	154
		3	165
		4	158
2	1	1	142
			...
27	7	3	186
		4	187
		1	266
		2	258
		3	255

Length: 861, dtype: int64

Figura 31: Distribucion de las clases

Los datos están organizados en función de:

- **Action ID:** Identificador de la acción realizada (1-27).
- **Person ID:** Identificador del sujeto (1-8)

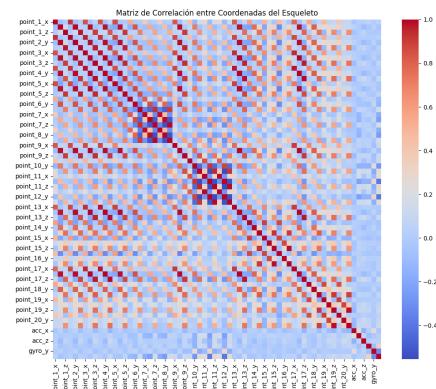
- **Test ID:** Número de repetición de la acción (1-4).

- **Valor numérico:** Cantidad de frames

Se observa variabilidad en los valores numéricos asociados a cada acción, lo que indica que la duración de las acciones no es uniforme. Por lo que un desbalance en la cantidad de muestras por acción podría afectar el rendimiento de los modelos de clasificación.

3.6. Correlación entre los atributos

La matriz de correlación muestra la relación entre las coordenadas del esqueleto en el conjunto de datos UTD-MHAD. Se observan fuertes correlaciones (en rojo oscuro) en coordenadas de la misma articulación y entre ejes relacionados, lo que indica dependencia entre ciertos movimientos. Las correlaciones más bajas (en azul) sugieren independencia entre algunas coordenadas y sensores iniciales.



3.7. Histograma de atributos

Se analiza la distribución de los atributos del esqueleto obtenidos a partir del conjunto de datos UTD-MHAD. Se han generado histogramas de cada coordenada (x,y, z) de los puntos clave del esqueleto con el objetivo de identificar patrones y tendencias en la distribución de los datos. La mayoría de las coordenadas siguen una distribución aproximadamente **gaussiana** o con ligeras asimetrías, lo que sugiere que los datos están bien capturados y reflejan posiciones naturales del esqueleto. Las coordenadas y tienen distribuciones con una mayor dispersión, indicando variabilidad en la altura de los puntos del cuerpo. Algunos histogramas presentan **picos secundarios**, lo que

podría indicar cambios bruscos de postura o errores en la captura de datos.

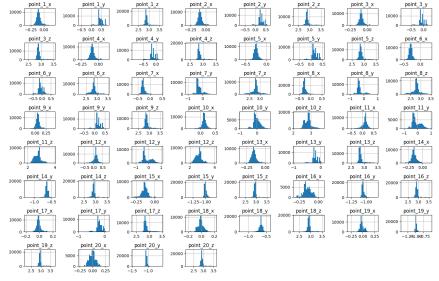


Figura 32: Histograma de puntos esqueleto

Por otro lado, los histogramas de las aceleraciones y las velocidades angulares, proporcionando información sobre la variabilidad de estos valores en las mediciones.

La asimetría en algunas distribuciones sugiere que ciertos movimientos pueden ser más frecuentes en una dirección específica. Se recomienda analizar las actividades individualmente para entender mejor estas diferencias.

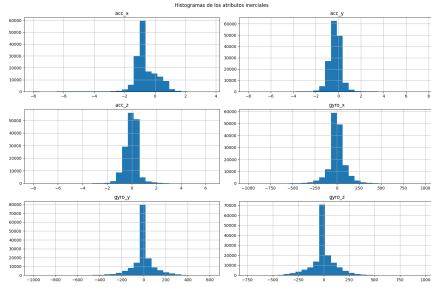


Figura 33: Histograma Inercial

3.8. Gráfica de densidad

Para los datos esqueletos, la mayoría de los puntos muestran distribuciones simétricas y cercanas a un valor medio, reflejando movimientos consistentes en esas coordenadas. Algunas distribuciones presentan asimetrías o múltiples picos, indicando variabilidad en la posición debido a cambios de postura o actividades diversas. Además, ciertos puntos clave tienen mayor dispersión, sugiriendo que segmentos como manos y pies son más móviles comparados con otros más estables como el torso o la cabeza.

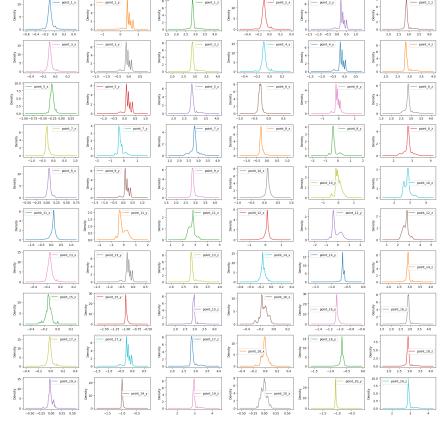


Figura 34: Densidad de datos esqueleto

El análisis de los datos iniciales sugiere que la distribución de las aceleraciones se encuentra centrada principalmente alrededor de valores cercanos a 0, lo que indica que la mayor parte de la actividad registrada se concentra en este rango. Sin embargo, se observa una ligera asimetría en algunos casos, lo que podría reflejar la presencia de patrones de movimiento no uniformes.

Además, se detectan valores atípicos negativos en las componentes *acc_x* y *acc_y*, los cuales podrían corresponder a movimientos bruscos o anomalías en los datos. Por otro lado, la densidad de valores se concentra en un rango estrecho, lo que refuerza la idea de que la mayor parte de la información se agrupa en torno a 0, sugiriendo una estabilidad relativa en los movimientos registrados. En cuanto a la componente *acc_z*, asociada a la aceleración vertical, se observa una menor dispersión en comparación con las otras direcciones. Esto podría indicar una menor variabilidad en la verticalidad del movimiento, lo que apunta a una mayor consistencia en esta dimensión. En conjunto, estos hallazgos permiten inferir que, aunque existen ciertas irregularidades y asimetrías, el comportamiento general de las aceleraciones tiende a mantenerse estable y centrado en valores próximos a 0.

Para el giroscopio presenta una alta dispersión, con colas largas que se extienden hasta valores extremos de aproximadamente ± 1500 . Esto indica la presencia de una amplia variabilidad en los valores registrados, con observaciones que se alejan significativamente del rango central. A diferencia de los datos de los acelerómetros, se observa una menor densidad de valores en la región central, lo

que sugiere que hay una menor concentración de datos en rangos bajos y, en cambio, una mayor dispersión hacia valores más grandes.

Además, se identifican posibles valores extremos (*outliers*) que podrían corresponder a mediciones atípicas o anomalías en los datos.

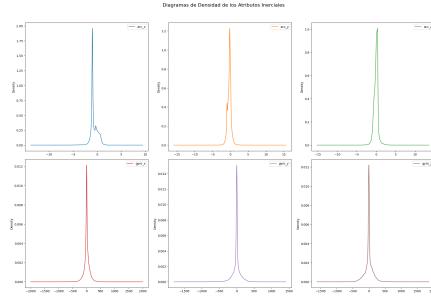


Figura 35: Densidad datos iniciales

3.9. Gráficas de caja y bigote

El siguiente análisis exploratorio consistió en la creación de boxplots (gráficas de cajas y bigotes) para los datos de puntos esqueleto e iniciales.

El análisis de los diagramas de caja y bigote de los atributos iniciales indica que la mayor parte de los datos se concentra en un rango estrecho, lo que sugiere una alta densidad de valores alrededor de la región central. Sin embargo, se observan valores atípicos extremos en ambos lados de la distribución, lo que indica la presencia de mediciones que se desvían significativamente del comportamiento general.

Los datos de los acelerómetros (*acc_x*, *acc_y*, *acc_z*) y giroscopios (*gyro_x*, *gyro_y*, *gyro_z*) muestran una distribución en la que la mayor parte de las observaciones se concentra en un rango estrecho, indicando que la mayoría de los valores se agrupan alrededor de una región central. Sin embargo, se observa la presencia de valores atípicos extremos en ambos lados de la distribución, lo que sugiere que existen mediciones que se desvían significativamente del comportamiento general.

En particular, se identifican numerosos *outliers* que se encuentran fuera de los bigotes en todos los atributos, lo que significa que estos puntos se alejan considerablemente del rango intercuartílico (IQR). Estos valores atípicos podrían estar asociados a movimientos bruscos, errores de medición o comportamientos anómalos en los datos.

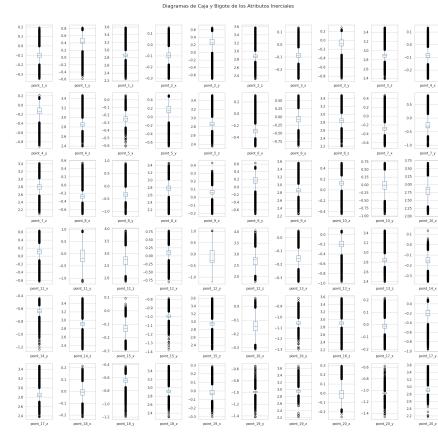


Figura 36: Esqueleto caja y bigote

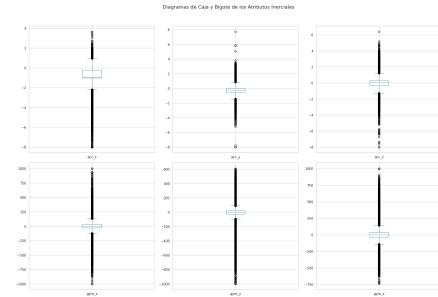


Figura 37: Enter Caption

3.10. Matriz de dispersión

Como última parte del análisis gráfico, se generó la matriz de dispersión de los atributos

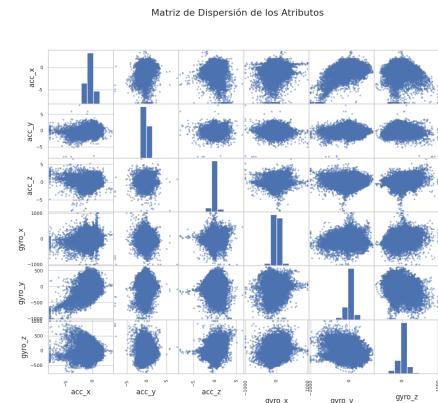
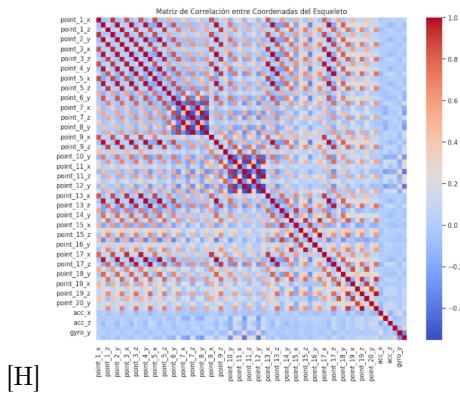


Figura 38

Se interpreta que los datos de aceleración tienen valores más concentrados, mientras que

los datos de giroscopio presentan una mayor dispersión, por lo que no hay una correlación lineal fuerte evidente entre la mayoría de las variables.

Imagen 1. Matriz de correlación entre atributos



En la imagen 1, Se observa correlaciones significativamente positivas entre coordenadas del mismo eje de distintos puntos articulares (point_1_x y point_3_x con $r > 0.85$). Esto sugiere un comportamiento coordinado entre múltiples articulaciones durante la ejecución de acciones, lo cual es coherente con la biomecánica de movimientos como caminar, saludar o girar el torso. Las correlaciones entre las coordenadas articulares y los datos iniciales fueron en general bajas ($r < 0.05$ en la mayoría de los casos), aunque no despreciables. Esto refleja que ambas modalidades capturan aspectos distintos del movimiento: mientras las coordenadas articulares representan posturas espaciales específicas, los sensores iniciales ofrecen información sobre aceleraciones y rotaciones globales, especialmente del brazo. La baja redundancia entre estas modalidades sugiere que su combinación puede ser beneficiosa para mejorar la discriminación entre clases de acción.

3.11. Escalamiento, normalización o estandarización

En cuanto a los tratamientos realizados, el creador de la base de datos menciona que hubo tres archivos corruptos; sin embargo, estos fueron eliminados previamente. Según el autor, todos los demás archivos se encuentran en perfectas condiciones. Con respecto a aplicar escalamiento, normalización o estandarización, en el caso de los datos y de las coordenadas esqueléticas, existe una

gran disparidad en magnitudes y unidades, lo que puede afectar negativamente a algoritmos sensibles a la escala, como máquinas de soporte vectorial o redes neuronales [5]. Además, se ha demostrado que la estandarización mejora la convergencia y estabilidad en técnicas de aprendizaje automático aplicadas al reconocimiento de actividades humanas, al reducir la varianza entre sujetos y condiciones de registro [6]. Esto es particularmente relevante en conjuntos multimodales como UTD-MHAD, donde las diferencias individuales y contextuales pueden introducir sesgos indeseados si no se normalizan las entradas [1]. Por lo tanto, estandarizar permite una representación más homogénea y comparativa de los patrones de movimiento. Por otra parte, la estrategia de segmentar los datos esqueléticos en ventanas temporales de 90 cuadros (equivalente a 3 segundos a 30 FPS) está respaldada por consideraciones computacionales. Esta duración es suficiente para capturar una instancia completa y representativa de la mayoría de las acciones humanas. De esta forma, una ventana de 3 segundos ofrece un equilibrio adecuado entre resolución temporal y contexto dinámico. Además, utilizar una segmentación por ventanas fijas permite estandarizar la entrada para los algoritmos de aprendizaje automático, facilitando la construcción de representaciones comparables entre acciones y sujetos [7]. Para los datos iniciales del conjunto UTD-MHAD, la frecuencia de muestreo de 50 Hz implica que se registran 50 muestras por segundo, lo que proporciona una alta resolución temporal útil para capturar variaciones rápidas del movimiento [8]. Sin embargo, cuando se pretende integrar estos datos con otros tipos de información sensorial, como los datos esqueléticos que operan a 30 FPS, es necesario armonizar las frecuencias de muestreo. En este caso, transformar los datos iniciales de 50 Hz a 30 Hz mediante un proceso de remuestreo controlado es una decisión justificada y común en aplicaciones multimodales. Esta transformación permite alinear temporalmente ambas modalidades, facilitando su fusión y reduciendo la complejidad del modelo sin una pérdida significativa de información relevante para la clasificación de acciones, dado que 30 Hz sigue siendo una tasa suficientemente alta para capturar la mayoría de los gestos humanos. En el procesamiento de datos esqueléticos para el reconocimiento de acciones humanas, la elección de atributos adecuados es clave pa-

ra capturar tanto la estructura como la dinámica del cuerpo humano. En este sentido, se emplearon tres tipos de características: angle_between, euclidean_distance y temporal_difference, por su eficacia comprobada en la literatura.

El atributo angle_between permite representar las relaciones angulares entre segmentos corporales, capturando la configuración postural del cuerpo de forma invariante a traslaciones y escalas. Esta representación angular es particularmente útil para distinguir actividades basadas en poses específicas [9]. La euclidean_distance entre pares de articulaciones seleccionadas ofrece una medida directa de la geometría del esqueleto, útil para capturar la extensión o contracción del cuerpo en ciertas acciones. Esta métrica es sencilla de calcular y se ha mostrado efectiva en múltiples estudios de reconocimiento basado en esqueletos [10]. Finalmente, la temporal_difference mide el cambio entre posiciones de articulaciones en frames sucesivos, lo que refleja indirectamente la velocidad del movimiento. Esto permite modelar la dinámica temporal de la acción, un aspecto crítico para diferenciar gestos similares con diferentes ritmos o intensidades [11]. En el análisis de datos iniciales provenientes de sensores como acelerómetros y giróscopos, el uso de atributos estadísticos es ampliamente aceptado por su capacidad para capturar patrones discriminativos de forma eficiente. La media y la desviación estándar son esenciales para representar la tendencia central y la variabilidad de la señal, lo cual es útil para distinguir acciones con diferentes niveles de intensidad o frecuencia [12] Por otro lado, el valor máximo y mínimo permiten identificar picos de aceleración o rotación, relevantes en movimientos abruptos o rápidos [13].

3.12. Validación Cruzada

Los resultados obtenidos a partir de la validación cruzada aplicada a los tres clasificadores (MLP, Random Forest y SVM) sobre las distintas representaciones de características permiten establecer varias observaciones relevantes en cuanto al desempeño de los modelos y la influencia de las técnicas de preprocesamiento utilizadas.

En primer lugar, se destaca de manera consistente que el rendimiento de todos los clasificadores mejora significativamente al utilizar el conjunto de características sin reducción de dimensionalidad. Specifically, the MLP clasificador alcanzó an accuracy of 84. 09 %, the most widely used of all

the experiments, conducted in a circle by Random Forest (83.51). %) y SVM (82.15 %). Este hallazgo sugiere que los métodos de reducción como PCA, Kernel PCA o SelectKBest podrían estar eliminando información discriminativa relevante para el aprendizaje de los modelos, al menos en el contexto específico de las características disponibles.

Por otro lado, al analizar las técnicas de reducción de dimensionalidad, Kernel PCA mostró un rendimiento ligeramente superior frente a PCA y SelectKBest, particularmente en el caso del MLP y SVM, lo cual indica que preservar relaciones no lineales en los datos puede aportar cierta ventaja frente a métodos lineales tradicionales. Sin embargo, aun así, los modelos con Kernel PCA no alcanzaron el nivel de desempeño observado sin reducción de características.

En cuanto a la comparación entre clasificadores, el perceptrón multicapa (MLP) demostró ser el más robusto y consistente entre las distintas representaciones de datos, manteniéndose como el mejor clasificador en términos de todas las métricas evaluadas (accuracy, precision, recall y F1-score). Esto sugiere que su capacidad para modelar relaciones no lineales complejas le otorga una ventaja significativa frente a Random Forest y SVM en esta tarea.

En resumen, los resultados indican que para este conjunto de datos y características, evitar la reducción de dimensionalidad proporciona mejores resultados, y que entre los clasificadores evaluados, el MLP es el más adecuado para generalizar correctamente sobre nuevas muestras. No obstante, estas conclusiones podrían variar en presencia de ruido o alta colinealidad entre características, por lo que es recomendable considerar un análisis de redundancia o pruebas adicionales sobre nuevas bases de datos antes de establecer una configuración definitiva para un sistema en producción.

3.13. Metodos de fusion

Los resultados obtenidos mediante la validación cruzada de los tres métodos de fusión —Random Forest como agregador, VotingClassifier como técnica de votación heterogénea, y AdaBoost como método de potenciación secuencial— revelan diferencias sustanciales en el desempeño de generalización, reflejando la sensibilidad de cada enfoque al tipo de representación de datos utilizada (en este caso, sin reducción de dimensionalidad) y a la forma en que combinan los clasificadores base.

Random Forest, que puede considerarse tanto un clasificador como un método de agregación basado en un conjunto de árboles de decisión, obtuvo el mejor rendimiento general, alcanzando un accuracy de 83.22 % y un F1-score de 82.63 %, superando a los otros métodos de fusión tanto en métricas globales como de precisión y recall. Este resultado reafirma la robustez de los bosques aleatorios cuando se aplican sobre conjuntos de características ricas, especialmente sin aplicar reducción de dimensionalidad, beneficiándose de su capacidad para manejar relaciones no lineales y alta dimensionalidad sin riesgo de sobreajuste inmediato.

El enfoque basado en votación suave entre clasificadores heterogéneos (regresión logística, bosque aleatorio y Naive Bayes) mostró un rendimiento ligeramente inferior, con un accuracy de 79.22 % y un F1-score de 78.70 %. Si bien esta técnica aporta diversidad en la toma de decisiones, los resultados sugieren que su eficacia podría estar limitada por la variabilidad en la capacidad de aprendizaje de sus componentes individuales, especialmente considerando que uno de ellos, Naive Bayes, asume independencia entre características, lo que rara vez se cumple en contextos reales como este.

Finalmente, AdaBoost mostró un desempeño marcadamente inferior, con un accuracy de apenas 25.95 %, lo que indica una alta inestabilidad en su capacidad para generalizar correctamente en este conjunto de datos. Este pobre rendimiento puede atribuirse a la sensibilidad de AdaBoost a ruido y a clasificadores base débiles que no logran aprender correctamente las complejidades del dominio. Es posible que el uso de características no reducidas haya exacerbado este efecto, al introducir redundancia o ruido que AdaBoost no pudo manejar adecuadamente.

En conclusión, los resultados evidencian que, para la base de datos considerada sin reducción de dimensionalidad, los métodos de agregación más estables como Random Forest y VotingClassifier son preferibles, mientras que AdaBoost, al menos en su configuración por defecto, no es una alternativa recomendable en este escenario. Estas conclusiones refuerzan la necesidad de evaluar cuidadosamente la compatibilidad entre los métodos de fusión y la estructura del conjunto de datos, así como de realizar un ajuste fino de hiperparámetros y selección de clasificadores base en escenarios donde se considere el uso de técnicas de potenciación

como AdaBoost.

4. PAMAP2 Dataset: Physical Activity Monitoring

4.1. Descripción del dataset

El conjunto de datos PAMAP2 contiene información correspondiente a 18 actividades físicas, de las cuales doce fueron registradas siguiendo un protocolo estandarizado, mientras que seis actividades adicionales fueron realizadas de forma opcional por algunos participantes. La base de datos fue construida a partir de la participación de nueve sujetos (ocho hombres y una mujer).

Durante la adquisición de datos, se utilizaron tres sensores iniciales (IMU) con una frecuencia de muestreo de 100 Hz, además de un pulsómetro con una frecuencia de 9 Hz. Los sensores IMU fueron posicionados de la siguiente manera:

- Uno en la muñeca del brazo dominante.
- Uno en el pecho.
- Uno en el tobillo del lado dominante.

Cada registro del conjunto de datos está compuesto por 54 columnas, distribuidas de la siguiente forma:

- 1 columna de marca temporal (timestamp, en segundos),
- 1 columna con el identificador de la actividad,
- 1 columna con la frecuencia cardíaca (bpm),
- 17 columnas correspondientes al IMU en la muñeca,
- 17 columnas correspondientes al IMU en el pecho,
- 17 columnas correspondientes al IMU en el tobillo.

Cada sensor IMU aporta las siguientes mediciones:

- Temperatura (°C),
- Acelerómetro 3D ($\pm 16g$),
- Acelerómetro 3D adicional ($\pm 6g$, resolución de 13 bits),
- Giroscopio 3D (rad/s),
- Magnetómetro 3D (μT),
- Orientación (cuaterniones).

Previo al análisis, se realizaron transformaciones iniciales en el conjunto de datos conforme a las

recomendaciones de su documentación. En particular, se eliminaron las siguientes columnas por considerarse irrelevantes para el análisis:

- Columnas de orientación de los sensores IMU, por carecer de validez en esta recopilación de datos.
- Columnas del acelerómetro 6g, debido a la falta de calibración respecto al acelerómetro principal y su propensión a saturarse ante aceleraciones superiores a 6g.

Asimismo, se excluyeron los registros asociados a la actividad con ID 0, ya que esta clase representa únicamente actividades de transición, como desplazamientos entre locaciones o tiempos de espera para preparación de equipos, sin valor analítico directo.

Respecto a los valores ausentes en los datos de los sensores IMU, se identificaron dos causas principales:

- Pérdidas puntuales debidas al uso de sensores inalámbricos, aunque ocurren de manera infrecuente.
- Fallas en la configuración del hardware, como interrupciones en la conexión o colapsos del sistema.

Estas situaciones generan valores ausentes aislados, es decir, rodeados por datos válidos dentro de la serie temporal. Para tales casos, se empleó interpolación lineal, ya que ha demostrado ser una técnica eficaz en este contexto [5].

En cuanto a la señal de frecuencia cardíaca, se observó una proporción elevada de datos nulos (alrededor del 90 %), atribuida a la baja tasa de muestreo del pulsómetro (9 Hz) en comparación con los sensores IMU (100 Hz). Aunque la literatura sugiere la imputación estadística en estos escenarios de observaciones faltantes [6], en este estudio se optó por descartar dicha variable, centrando el análisis exclusivamente en los datos generados por los sensores iniciales.

4.2. Experimentación

4.2.1. Características y métodos de reducción de dimensionalidad

Se optó por las características descritas en [8] con una ventana de tiempo de 3s ya que nos menciona que para datos iniciales son las que proveen mejores resultados.

Así mismo para las técnicas de reducción de dimensionalidad se optaron por las técnicas de análisis de componente principal, análisis de componentes principales de kernel y la técnica de selección de características por medio de la prueba ANOVA.

4.2.2. Modelos de Aprendizaje y Métodos de fusión

Para la evaluación de los modelos individuales se optaron por:

- Perceptron multicapa (MLP).
- Regresión Logística (RL).
- Bosques Aleatorios (RF).

Para los métodos de fusión se optó por:

- Random Forest Aggregation (RFA).
- Votación con los clasificadores de regresión logística, bosques aleatorios y bayes ingenuo (Voto).
- AdaBoost con arboles de máxima profundidad 2 (AdBo)

4.2.3. Métricas para medir el desempeño

Las métricas seleccionadas para emplear fueron F1-Score, Accuracy, Precision y Recall.

4.2.4. Resultados

En las figuras 39 y 40 se puede apreciar los resultados obtenidos con los modelos individuales con el dataset de características original y los resultantes de cada técnica de reducción de características:

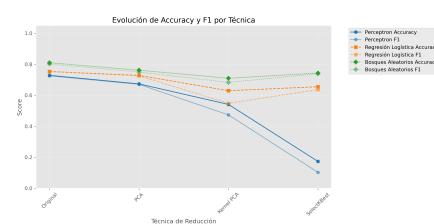


Figura 39: Evolución del desempeño de los modelos con las transformaciones

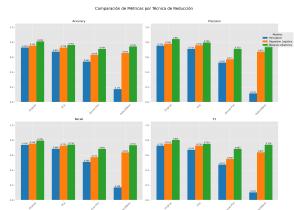


Figura 40: Comparacion Desempeño Modelos

En las figuras 41 y 42 se puede apreciar los resultados obtenidos con los métodos de fusión con el dataset de características original y los resultantes de cada técnica de reducción de características:

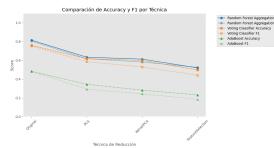


Figura 41: Evolucion del desempeño metodos de fusión

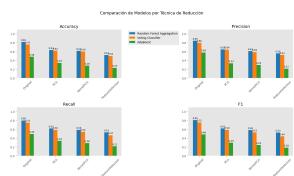


Figura 42: Comparación desempeño Metodos de Fusion

4.3. Discusión

Para los modelos individuales el análisis comparativo muestra que el mejor desempeño se obtuvo utilizando el conjunto de características original, donde el modelo de Bosques Aleatorios alcanzó una precisión y F1 superiores, demostrando su capacidad para manejar la complejidad del dataset sin necesidad de reducción. Al aplicar PCA, Kernel PCA y SelectKBest, el rendimiento disminuyó en mayor o menor medida para todos los modelos, destacando una pérdida significativa especialmente con Kernel PCA y un desempeño muy pobre con SelectKBest, particularmente en el Perceptrón. Aunque PCA logró resultados aceptables, también implicó una pérdida de información relevante. En general, las técnicas de reducción no aportaron mejoras y, en la mayoría de los casos,

perjudicaron la capacidad predictiva de los modelos. Por lo tanto, se recomienda mantener el conjunto completo de características, salvo que existan limitaciones computacionales o se busque interpretabilidad.

Por otro lado, en el caso de los métodos de fusión se pueden extraer varias conclusiones relevantes. En primer lugar, **Random Forest Aggregation** se posiciona como el método de ensamble más sólido, obteniendo el mejor rendimiento con el conjunto original de características (Accuracy: 0.815, F1: 0.804), superando incluso al modelo Random Forest individual. Esto sugiere que la agregación mejora la estabilidad y capacidad predictiva del modelo, beneficiándose del uso del conjunto completo de información sin necesidad de reducción dimensional. Por su parte, **Voting Classifier** también muestra un rendimiento aceptable con el dataset original, aunque inferior al de Random Forest, lo cual es esperable dada la naturaleza más simple del mecanismo de votación. En contraste, **AdaBoost** tuvo un rendimiento muy pobre en todas las configuraciones, con una caída especialmente marcada al aplicar técnicas de reducción como PCA, Kernel PCA o selección de características. Esto podría deberse a que AdaBoost es muy sensible al ruido o a características no informativas, y en este caso, la reducción de dimensionalidad no logró preservar un conjunto útil para su funcionamiento.

Al aplicar PCA y Kernel PCA, todos los métodos de ensamble experimentaron una disminución general en el rendimiento, lo que refuerza la idea de que estas transformaciones, si bien pueden ser útiles para simplificar el espacio de características, también pueden eliminar información clave para la clasificación. En particular, Voting y AdaBoost se vieron más afectados, mientras que Random Forest logró mantener un desempeño moderado. Finalmente, al emplear **SelectKBest**, se observaron los peores resultados globales, especialmente con AdaBoost, indicando que esta técnica de selección no logró conservar las características más relevantes para los clasificadores de ensamblado.

5. Conclusiones

Este trabajo evaluó el impacto de distintas técnicas de reducción de características (PCA, Kernel PCA y ANOVA F-test) y métodos de fusión (Voting, Boosting, Random Forest) sobre el

desempeño de modelos de clasificación (Random Forest, Logistic Regression, Perceptron, MLP y SVM) en tres datasets: MHealth, UTD-MHAD y PAMAP2.

En general, el modelo Random Forest demostró ser el más robusto y preciso, especialmente en escenarios con alta dimensionalidad. El Perceptron mostró un rendimiento competitivo, superando en algunos casos a la regresión logística, siendo el más efectivo en el dataset UTD.

La técnica ANOVA F-test fue la técnica más eficaz al permitir una reducción significativa del número de atributos (de 164 a 50 en el caso de MHealth), mejorando o manteniendo el rendimiento de los modelos, especialmente Random Forest y Logistic Regression. PCA y Kernel PCA tuvieron efectos mixtos: sólo ayudaron ligeramente en modelos lineales como la regresión logística, y en algunos casos empeoraron el rendimiento de otros modelos.

El VotingClassifier y Random Forest como método de ensamblado ofrecieron mejoras notables en estabilidad y rendimiento general. El Boosting (AdaBoost) mostró resultados inconsistentes, particularmente en el dataset UTD, donde fue muy sensible al ruido y datos no reducidos.

En todos los datasets, evitar la reducción de dimensionalidad tendió a ofrecer mejores resultados, a menos que se aplicara una técnica robusta como ANOVA F-test. Los datos multimodales (inerciales + esqueléticos) del dataset UTD confirmaron que la combinación de sensores heterogéneos mejora la capacidad de discriminación entre clases si se manejan adecuadamente.

Referencias

- [1] Ciencia de Datos en R. (2024). *Capítulo 26: Clasificador k-vecinos más próximos*. Disponible en: <https://cdr-book.github.io/cap-knn.html>. [Último acceso: 5 de marzo de 2025].
- [2] Akoglu H. (2018). *User's guide to correlation coefficients*. National Library of Medicine. Disponible en: <https://pmc.ncbi.nlm.nih.gov/articles/PMC6107969/>. [Último acceso: 6 de marzo de 2025].
- [3] Liu, J., Yao, J., Zhou, Q., Wang, Z., & Huang, L. (2023). *LSTMAE-DWSSLM: A unified approach for imbalanced time series data classification*. Applied Intelligence, 53, 21077–21091. Disponible en: <https://link.springer.com/article/10.1007/s10489-023-04642-0>.
- [4] Reiss, A., Hendeby, G., & Stricker, D. (2014). *A novel confidence-based multiclass boosting algorithm for mobile physical activity monitoring*. Personal and Ubiquitous Computing, 19(1), 105–121. <https://doi.org/10.1007/s00779-014-0816-x>.
- [5] Tawakuli, A., Kaiser, D., & Engel, T. (2023). *Experience: differentiating between isolated and sequence missing data*. Journal of Data and Information Quality, 15(2), 1–15.
- [6] Allison, P. (2002). *Missing data*. In SAGE Publications, Inc. eBooks. <https://doi.org/10.4135/9781412985079>
- [7] Dimitris, E. (s.f.). *Normalization and Standardization*. Disponible en: <https://deffro.github.io/tutorials/normalization-standardization/> [Último acceso: 6 de marzo de 2025]
- Vinay, S. (2021). *Standardization in Machine Learning*.
- [8] Garcia-Ceja E., Galván-Tejada C., Brená R. (2018). *Multi-view stacking for activity recognition with sound and accelerometer data* Disponible en: <https://www.sciencedirect.com/science/article/abs/pii/S1566253516301932?via%3Dihub> [Último acceso: 6 de marzo de 2025]
- [9] Afzali Arani, M. S., Costa, D. E., & Shihab, E. (2021). Human Activity Recognition: A Comparative Study to Assess the Contribution Level of Accelerometer, ECG, and PPG Signals. Sensors, 21(21), 6997. Recuperado de: <https://www.mdpi.com/1424-8220/21/21/6997> [Último acceso: 6 de marzo de 2025]
- [10] Camps, V., & Harle, R. (2022). Are Gyroscopes an Added Value in Leave-One-Subject-Out Activity Recognition with IMUs? En 2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops) (pp. 1–6). IEEE. Recuperado de: <https://ieeexplore.ieee.org/document/9871845>

- [11] Müller, A. & Guido S., (2017). Introduction to Machine Learning with Python. O'Reilly Media, Inc. ISBN: 9781449369415.
- [12] Reiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. Recuperado de: <https://link.springer.com/article/10.1023/A:1010933404324>
- [13] Zaki, Z., Shah, M. A., Wakil, K., & Sher, F. (2020). Logistic Regression Based Human Activities Recognition. Recuperado de: https://www.researchgate.net/publication/341106657_LOGISTIC_REGRESSION_BASED_HUMAN_ACTIVITIES_RECOGNITION
- [14] Bulling, A., Blanke, U., & Schiele, B. (2014). A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys*, 46(3), 1–33. <https://doi.org/10.1145/2499621>
- [15] Nguyen, D.-A., & Le-Khac, N.-A. (2024). SoK: Behind the Accuracy of Complex Human Activity Recognition Using Deep Learning. arXiv preprint arXiv:2405.00712. Recuperado de: <https://arxiv.org/html/2405.00712v2#bib.bib18>
- [16] Chen, C., Jafari, R., & Kehtarnavaz, N. (2015). UTD-MHAD: A Multimodal Dataset for Human Action Recognition Utilizing a Depth Camera and a Wearable Inertial Sensor. En *2015 IEEE International Conference on Image Processing (ICIP)* (pp. 168–172), Quebec City, QC, Canadá. doi: <https://doi.org/10.1109/ICIP.2015.7350781>
- [17] Microsoft. Kinect for Windows. Recuperado de: <http://www.microsoft.com/en-us/kinectforwindows/>
- [18] Chen, C., Liu, K., Jafari, R., & Kehtarnavaz, N. (2014). Home-based senior fitness test measurement system using collaborative inertial and depth sensors. En *Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Chicago, IL, pp. 4135–4138.
- [19] Chen, Y., Yang, J., Wang, S., & Dong, H. (2015). Sensor-based activity recognition using deep learning: A survey. *Pattern Recognition Letters*, 73*, 3–12. <https://doi.org/10.1016/j.patrec.2015.04.019>
- [20] Ronao, C. A., & Cho, S. B. (2016). Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Systems with Applications*, 59*, 235–244.
- [21] Ordóñez, F. J., & Roggen, D. (2016). Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16*(1), 115.
- [22] Lara, O. D., & Labrador, M. A. (2013). A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys & Tutorials*, 15*(3), 1192–1209.
- [23] Wang, J., Liu, Z., Wu, Y., & Yuan, J. (2012). Mining actionlet ensemble for action recognition with depth cameras. En *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1290–1297.
- [24] Vemulapalli, R., Arrate, F., & Chellappa, R. (2014). Human action recognition by representing 3D skeletons as points in a Lie group. En *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 588–595.
- [25] Offli, F., Chaudhry, R., Kurillo, G., Vidal, R., & Bajcsy, R. (2014). Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition. *Journal of Visual Communication and Image Representation*, 25*(1), 24–38.
- [26] Banos, O., Galvez, J. M., Damas, M., Pomares, H., & Rojas, I. (2014). Window size impact in human activity recognition. *Sensors*, 14*(4), 6474–6499.
- [27] Kwapisz, J. R., Weiss, G. M., & Moore, S. A. (2011). Activity recognition using cell phone accelerometers. *SIGKDD Explorations*, 12*(2), 74–82.
- [28] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182.
- [29] Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). Springer Series in Statistics. Springer. Recuperado de: <https://link.springer.com/book/10.1007/b98835>

- [30] Preece, S. J., Goulermas, J. Y., Kenney, L. P. J., Howard, D., Meijer, K., & Crompton, R. (2009). A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. *IEEE Transactions on Biomedical Engineering*, 56(3), 871–879. <https://doi.org/10.1109/TBME.2008.2006190>
- [31] Jadama, Ansumana & Toray, Modou. (2024). Ensemble Learning: Methods, Techniques, Application. 10.13140/RG.2.2.28017.08802. Recuperado de: https://www.researchgate.net/publication/381773312_Ensemble_Learning_Methods_Techniques_Application
- [32] Almaslukh, B., AlMuhtadi, J., & Artooli, A. (2018). An effective deep auto-encoder approach for online smartphone-based human activity recognition. *Procedia Computer Science*, 141, 88–95. Recuperado de: <https://www.sciencedirect.com/science/article/pii/S1877050918319719>
- [33] M. Shoaib, et al. (2016). Human Activity Recognition: A Comparative Study to Assess the Contribution Level of Accelerometer, ECG, and PPG Signals.
- [34] Schölkopf, B., Smola, A., & Müller, K.-R. (1997). Kernel principal component analysis. En W. Gerstner, A. Germond, M. Hasler, & J.-D. Nicoud (Eds.), *Artificial Neural Networks — ICANN 1997* (pp. 583–588). Springer. <https://doi.org/10.1007/BFb0020217>
- [35] Mika, S., Rätsch, G., Weston, J., Schölkopf, B., & Müller, K.-R. (1999). Fisher Discriminant Analysis with Kernels. En *Proceedings of the 1999 9th IEEE Workshop on Neural Networks for Signal Processing (NNSP'99)* (pp. 41–48). Madison, WI, USA: IEEE.
- [36] MHEALTH Dataset - <https://archive.ics.uci.edu/dataset/319/mhealth+dataset>
- [37] Reiss, A., & Stricker, D. (2012). Introducing a new benchmarked dataset for activity monitoring.
- [38] Kira, K., & Rendell, L. (1992). The Feature Selection Problem: Traditional Methods and a New Algorithm.
- [39] Zhou, Z. H. (2012). *Ensemble Methods: Foundations and Algorithms*.
- [40] Hammerla, N. Y., Halloran, S., & Plötz, T. (2016). Deep, Convolutional, and Recurrent Models for Human Activity Recognition Using Wearables. *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, 1533–1540. <https://arxiv.org/abs/1604.08880>
- [41] Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting.