Informe Detallado: Proyecto de Clasificación de Actividades Físicas Basadas en Datos Sensoriales

Informe final del proyecto que incluye todas las etapas realizadas de Elida Peña:

---

1. Preprocesamiento de Datos

Descripción de los datos

- Los datos contienen mediciones sensoriales capturadas por dispositivos durante la realización de actividades físicas. Las variables independientes corresponden a sensores (e.g., acelerómetros y giroscopios) y la variable dependiente representa el tipo de actividad (e.g., "WALKING", "LAYING").

Pasos realizados:

1. **Normalización:**
    - Los datos de los sensores fueron escalados utilizando el método `StandardScaler` para garantizar que todas las variables tuvieran igual peso en el análisis.
2. **Codificación de etiquetas:**
    - Las actividades fueron transformadas en un formato categórico mediante `LabelEncoder` y `to_categorical` para ser procesadas en el modelo.
3. **División del conjunto de datos:**
    - El dataset fue dividido en 70% para entrenamiento, 15% para validación y 15% para prueba.

---

2. Análisis No Supervisado

Aplicación de PCA

- Se realizó un análisis de componentes principales (PCA) para reducir la dimensionalidad y analizar la estructura de los datos.

Resultados:

1. Se identificaron 73 componentes principales necesarios para explicar el 95% de la varianza total.
2. Se graficaron los datos reducidos a dos componentes principales para visualizar las actividades físicas.
3. Se analizaron las distribuciones de las actividades utilizando diferentes métodos de escalado (e.g., sin escalado, Min-Max, StandardScaler y RobustScaler).

Conclusión:

- El PCA proporcionó una representación visual clara de las actividades físicas, mostrando cómo se agrupan según patrones comunes en los sensores. Este análisis fue útil para entender la separabilidad de las actividades antes de construir el modelo MLP.

---

3. Modelado MLP

Configuración del modelo

1. Arquitectura:

    - ✓ **Entrada:** Dimensión igual al número de características sensoriales.
    - ✓ **Capas ocultas:**

    - 1ª capa: 128 neuronas, función de activación ReLU, dropout 30%.
    - 2ª capa: 64 neuronas, función de activación ReLU, dropout 30%.

    - ✓ **Capa de salida:** Número de clases igual a las actividades, función de activación softmax.

2. **Optimización:**

    - ✓ Optimizador: `Adam` con tasa de aprendizaje de 0.001.
    - ✓ Pérdida: `categorical_crossentropy`.

Entrenamiento:

- El modelo fue entrenado durante 50 épocas con un tamaño de batch de 32, utilizando validación cruzada.

---

4. Evaluación del Modelo

Resultados en el conjunto de prueba

1. **Métricas generales:**

    - ✓ **Precisión:** 97.97%.
    - ✓ **Pérdida:** 0.0871.

2. **Reporte de Clasificación:**

    - ✓ Las actividades como "LAYING" y "WALKING_UPSTAIRS" obtuvieron F1-score perfecto (1.00).
    - ✓ Actividades similares como "SITTING" y "STANDING" presentaron ligeras confusiones con F1-scores de 0.95.

3. **Matriz de Confusión:**

    - ✓ La matriz mostró que el modelo tiene un alto grado de precisión en todas las clases, aunque con pequeñas confusiones entre actividades similares.

---

5. Visualización de Curvas de Aprendizaje

Pérdida:

- Tanto la pérdida en entrenamiento como en validación disminuyen rápidamente en las primeras épocas y se estabilizan en valores bajos, indicando que el modelo aprendió eficientemente sin sobreajustarse.

Precisión:

- La precisión en entrenamiento y validación es alta desde las primeras épocas, estabilizándose alrededor del 97-98%.

---

6. Conclusiones

1. **Insights del Análisis No Supervisado:**

    - ✓ El PCA demostró que los datos sensoriales contienen patrones claros que separan las actividades físicas.
    - ✓ Las características más importantes son aquellas que capturan patrones específicos de movimiento (e.g., aceleración y rotación).

2. **Rendimiento del MLP:**

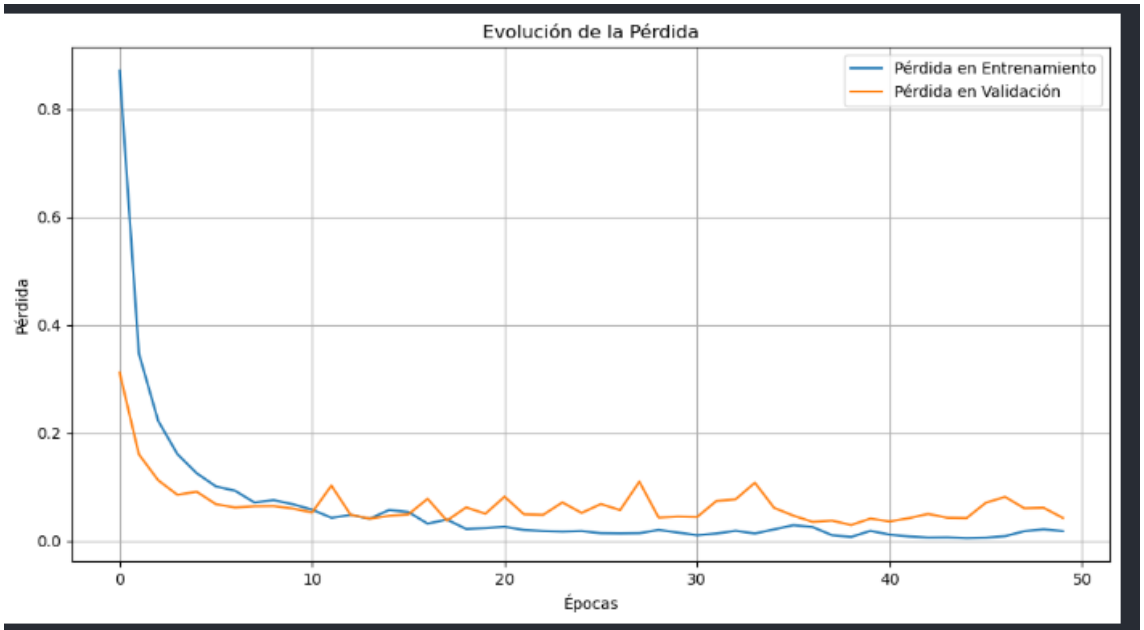
    - ✓ El modelo MLP es altamente efectivo para predecir actividades físicas con datos sensoriales, logrando métricas cercanas a la perfección.
    - ✓ Sin embargo, actividades similares como "SITTING" y "STANDING" requieren un análisis más profundo debido a ligeras confusiones.

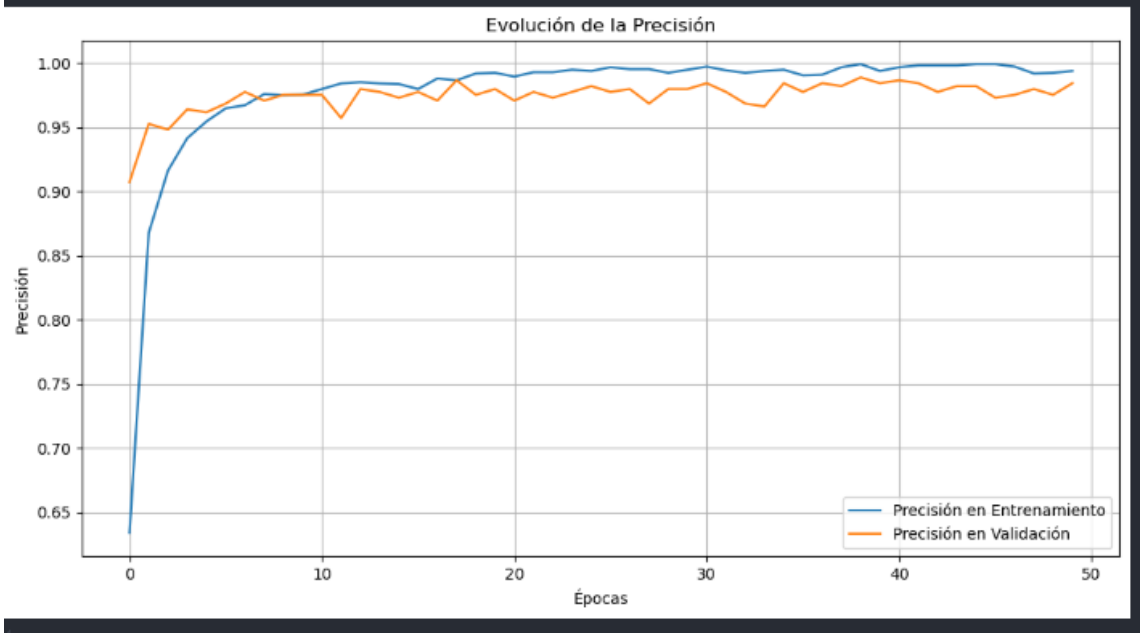
Anexos

Se incluyen las siguientes visualizaciones:

1. Gráfico de varianza explicada acumulada por PCA.
2. Gráficos de dispersión de los datos reducidos por PCA.
3. Reporte de clasificación y matriz de confusión del modelo.
4. Curvas de aprendizaje de pérdida y precisión.

Visualizaciones que demuestren los resultados obtenidos.





Interpretación del gráfico Evolucion de Perdida:

Comportamiento inicial:

La línea azul (pérdida en entrenamiento) comienza en un valor alto (~0.8), indicando que el modelo inicialmente tiene un desempeño deficiente sobre el conjunto de entrenamiento. La línea naranja (pérdida en validación) también comienza relativamente alta (~0.2-0.4), aunque más baja que la pérdida en entrenamiento. Convergencia rápida:

Ambas curvas muestran una rápida disminución durante las primeras épocas (aproximadamente las primeras 10). Esto indica que el modelo está aprendiendo de manera efectiva, reduciendo el error en ambos conjuntos. Estabilidad:

Después de las primeras 10 épocas, las curvas se estabilizan. La pérdida en entrenamiento (azul) se mantiene baja y consistente. La pérdida en validación (naranja) también permanece baja y relativamente estable, con ligeras fluctuaciones hacia el final (posiblemente por ruido o una ligera sobre-adaptación al conjunto de validación). Conclusión: El modelo parece haberse entrenado de manera efectiva, logrando una pérdida baja tanto en entrenamiento como en validación. Esto implica que el modelo está aprendiendo patrones relevantes de los datos sin memorizar en exceso, siendo capaz de generalizar bien al conjunto de validación.

Interpretación del grafico de Evolución de la Precisión:

Comportamiento inicial:

La línea azul (precisión en entrenamiento) comienza baja (~0.65), lo que indica que el modelo inicialmente tiene un desempeño limitado sobre el conjunto de entrenamiento. La línea naranja (precisión en validación) comienza más alta (~0.90), lo que sugiere que incluso desde el inicio el modelo logra generalizar razonablemente bien a los datos de validación. Rápido incremento inicial:

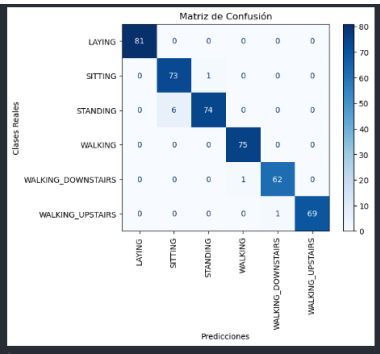
Durante las primeras 10 épocas, ambas curvas aumentan rápidamente. Esto refleja que el modelo está aprendiendo patrones importantes en los datos. La precisión en entrenamiento supera la precisión en validación durante las primeras épocas, lo cual es típico durante el proceso de ajuste. Estabilización:

Después de aproximadamente 20 épocas, ambas curvas se estabilizan cerca del valor máximo (~0.98-0.99). La precisión en validación se mantiene ligeramente por debajo de la precisión en entrenamiento, pero ambas están muy cercanas, lo que indica que el modelo tiene un buen nivel de generalización y no muestra signos significativos de sobreajuste (overfitting). Resultados finales:

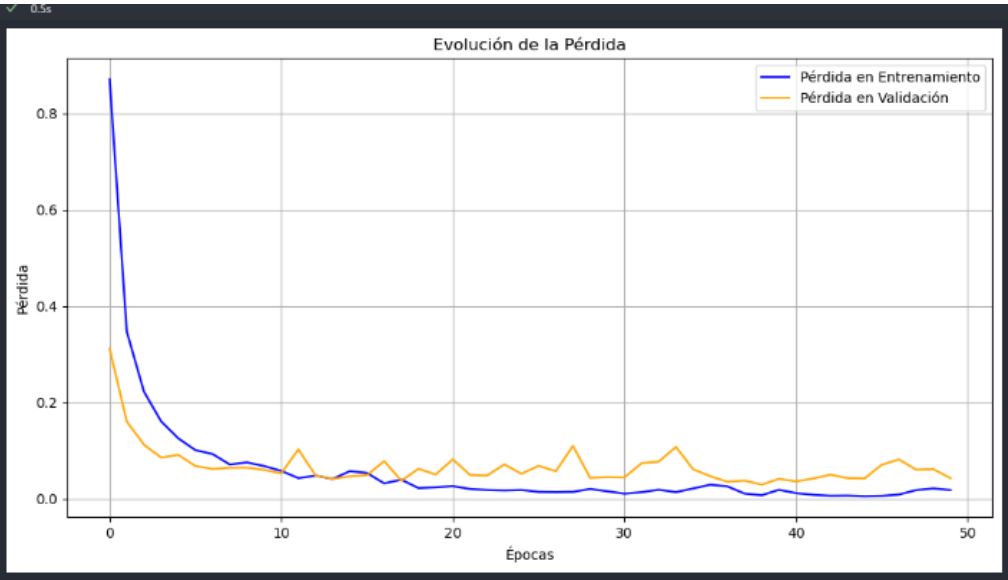
La precisión en entrenamiento al final alcanza un valor cercano al 100% (1.0), indicando que el modelo se adapta muy bien a los datos de entrenamiento. La precisión en validación permanece alta (~0.98-0.99), lo que sugiere que el modelo tiene un excelente desempeño sobre datos que no ha visto antes, lo cual es una señal de buena capacidad predictiva. Conclusión: El modelo logra una precisión muy alta tanto en entrenamiento como en validación, con diferencias mínimas entre ambas curvas. Esto implica que el modelo es capaz de aprender patrones relevantes en los datos sin caer en sobreajuste, y está bien generalizado para realizar predicciones sobre nuevos datos.

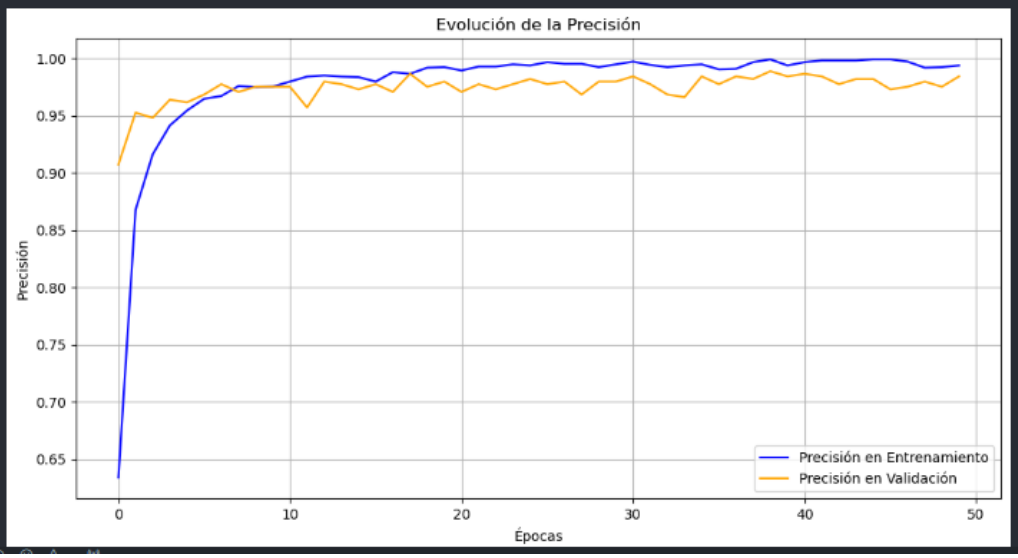
4. Evaluación del modelo:

- Evalúa el rendimiento del MLP utilizando métricas como la precisión, el recall, el F1-score y la matriz de confusión. Discute si el modelo es capaz de predecir correctamente las distintas actividades físicas.
- Visualiza las curvas de aprendizaje, mostrando cómo evoluciona la pérdida y la precisión a lo largo del entrenamiento.



- **Conclusión:** El modelo es muy capaz de predecir correctamente las actividades físicas basándose en los datos sensoriales. Con un F1-score promedio de 0.98 y una precisión en el conjunto de prueba cercana al 98%, demuestra ser confiable para la clasificación.

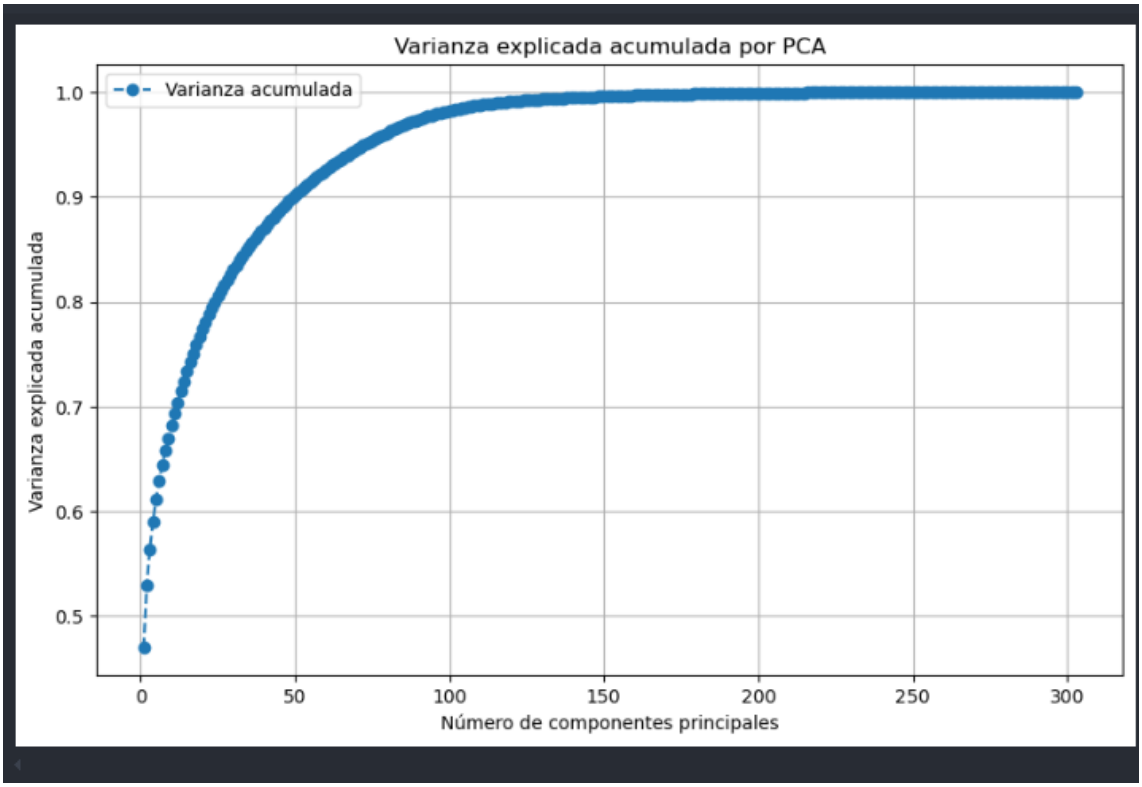




Evolución General:

Las curvas son consistentes antes y después, pero los gráficos actuales muestran un modelo bien ajustado que no ha sufrido ni underfitting ni overfitting. Las pequeñas oscilaciones en las últimas épocas son normales y reflejan variaciones menores debido al conjunto de datos.

Evaluación Final: El modelo ha evolucionado correctamente, manteniendo una baja pérdida y una alta precisión, lo cual confirma su capacidad para predecir de manera confiable las actividades físicas en el conjunto de validación.



Interpretación:

Inicio rápido: Los primeros componentes principales explican la mayor parte de la información de los datos. Por ejemplo, los primeros 50 componentes ya representan casi el 90% de la variabilidad total. Esto significa que mucha información se concentra en pocas dimensiones.

Curva plana hacia el final:

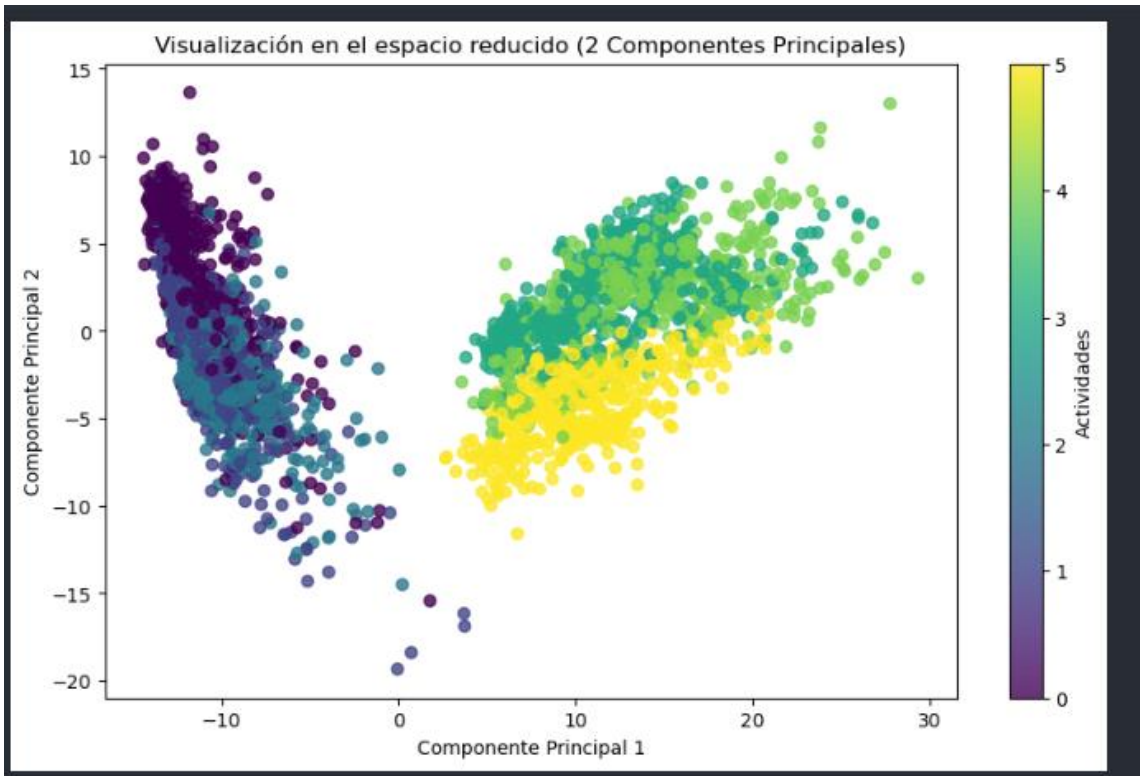
A partir de cierto punto, agregar más componentes no añade mucha información nueva. En este caso, después de unos 73 componentes, la curva casi se aplana. Esto nos dice que esos componentes adicionales no son tan útiles.

Número clave de componentes:

Si elegimos 73 componentes principales, podemos explicar el 95% de la información de los datos originales. Esto es un gran logro porque reducimos la complejidad del problema sin perder demasiada información.

En resumen:

Este gráfico nos muestra que PCA es muy útil para reducir la cantidad de datos con los que trabajamos (de cientos de dimensiones a unas pocas), manteniendo casi toda la información importante. Esto hace que cualquier análisis o modelo posterior sea más rápido y eficiente.



Interpretación:

1. **Agrupamientos claros:**

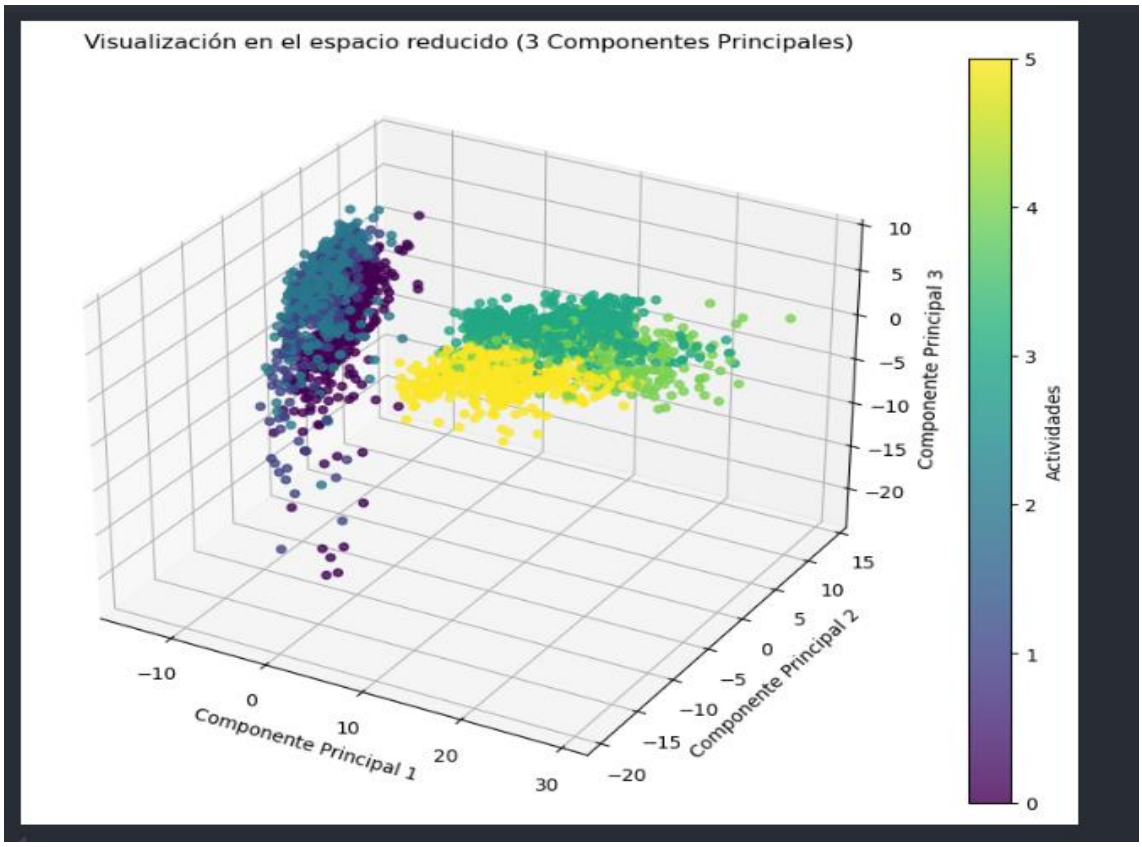
   - ✓ Las actividades tienden a formar grupos distintos en este espacio reducido. Por ejemplo, los puntos de color púrpura (una actividad específica) están separados de los puntos amarillos o verdes, lo que sugiere que PCA puede separar bien estas actividades.

2. **Relación entre actividades:**

   - ✓ Algunas actividades están más cercanas en el espacio, como los puntos verdes y amarillos, lo que podría indicar que estas actividades tienen características similares en los datos originales.

3. **Reducción exitosa:**

   - ✓ Aunque la información se comprimió de muchas dimensiones a dos, la separación entre grupos sugiere que los componentes principales capturaron bien las diferencias más importantes entre las actividades.



Interpretación:

1. **Separación de las actividades:**

   - ✓ Las actividades se agrupan en el espacio tridimensional según patrones similares en sus datos originales. Esto indica que el PCA logró preservar diferencias clave entre actividades.
   - ✓ Por ejemplo, las actividades representadas por colores púrpura y amarillo están bien separadas, lo que sugiere que son fáciles de distinguir.

2. **Dimensiones adicionales:**

   - ✓ A diferencia del gráfico 2D, esta representación con tres componentes principales permite observar una separación más clara en casos donde las actividades podían solaparse en el plano 2D. Esto agrega profundidad a la interpretación de los datos.

3. **Utilidad del PCA:**

   - ✓ El gráfico destaca que los tres componentes principales explican gran parte de la variabilidad en los datos. Esto facilita visualizar diferencias entre actividades sin usar todas las dimensiones originales.

En resumen, el gráfico evidencia que el PCA no solo simplifica los datos, sino que también resalta patrones importantes que ayudan a distinguir actividades físicas, lo cual es útil para análisis y modelos predictivos.