Predicción de Actividades Físicas con MLP y PCA

Introducción

Este informe presenta los resultados obtenidos mediante el análisis y modelado de datos sensoriales para la predicción de actividades físicas utilizando técnicas de aprendizaje automático. En este trabajo, se emplean dos enfoques principales: un análisis no supervisado utilizando PCA (Análisis de Componentes Principales) y un modelo MLP (Perceptrón Multicapa) con y sin Dropout. El objetivo es predecir actividades físicas como caminar, estar de pie, sentarse, etc., basándose en lecturas de aceleración en los ejes X, Y y Z, proporcionadas por sensores de smartphones.

Datos Utilizados:

El conjunto de datos utilizado en este análisis proviene de la *Human Activity Recognition with Smartphones Dataset*, disponible en *Kaggle*. Este *dataset* contiene mediciones de aceleración en los tres ejes (X, Y, Z) tomadas por sensores en un smartphone mientras los usuarios realizan diferentes actividades físicas, como caminar, estar sentado, subir escaleras, entre otras.

Objetivos del Informe:

El propósito de este informe es proporcionar un análisis detallado de las siguientes etapas:

1. Preprocesamiento de los datos, que incluye la carga y limpieza de los datos.
2. Análisis no supervisado mediante PCA, para reducir la dimensionalidad del conjunto de datos y visualizar las relaciones entre las actividades físicas.
3. Desarrollo de un modelo MLP para predecir las actividades físicas.
4. Evaluación del rendimiento de los modelos y su capacidad de generalización a nuevos datos.

Metodología:

El análisis no supervisado (PCA) permite una representación más eficiente de los datos, mientras que el modelo MLP se utiliza para predecir las clases de actividades. Además, se compara el rendimiento de ambos modelos, con y sin *Dropout*, para evaluar la capacidad de generalización y evitar el sobreajuste.

Preparado por: Andrea Echague Morel

1. Preprocesamiento de Datos

Carga y Exploración de los Datos:

El conjunto de datos se cargó y se exploró para identificar las variables relevantes para la clasificación de actividades físicas. El conjunto incluye lecturas de aceleración en los ejes X, Y y Z, que son fundamentales para predecir actividades como caminar, estar de pie, sentarse, etc.

Manejo de Valores Nulos y Normalización:

Se verificaron los valores nulos en el conjunto de datos. En caso de encontrarse valores faltantes, estos se imputaron utilizando la media de las respectivas columnas. Luego, las características sensoriales fueron normalizadas utilizando *StandardScaler*. Este paso es crucial, ya que asegura que todas las características tengan la misma escala, evitando que algunas dominen a otras debido a diferencias en magnitudes.

División de Datos:

El conjunto de datos fue dividido en conjuntos de entrenamiento y prueba. La división se realizó de manera que todas las actividades estuvieran representadas en ambos conjuntos, asegurando que los modelos entrenaran con una muestra representativa de las clases y evaluaran su rendimiento en datos no vistos.

2. Análisis No Supervisado (PCA)

Aplicación de PCA:

Se aplicó el Análisis de Componentes Principales (PCA) para reducir la dimensionalidad del conjunto de datos y visualizar las relaciones entre las actividades físicas. PCA permite identificar las componentes principales que capturan la mayor parte de la variabilidad de los datos.

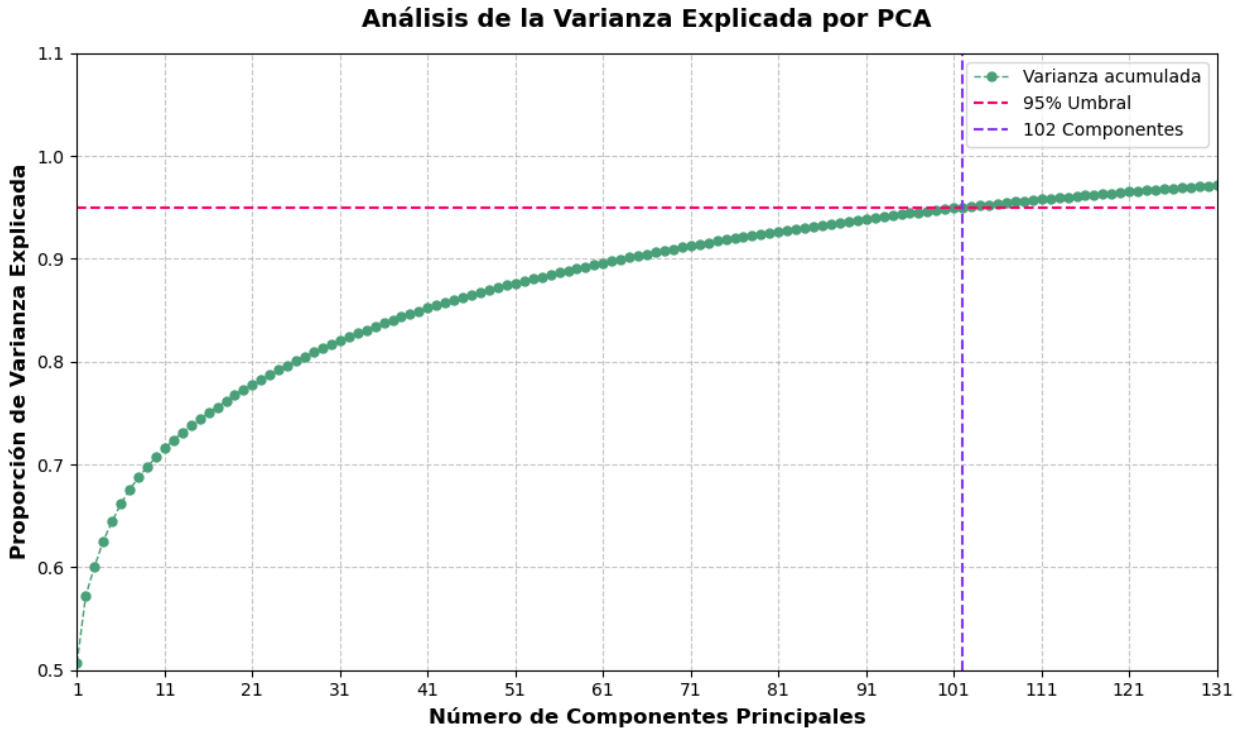
Resultados del PCA:

- Primera Componente Principal: Explicó 50.69% de la varianza total en los datos.
- Segunda Componente Principal: Explicó un 6.57% adicional de la varianza.
- Juntas, estas dos componentes explican 57.26% de la variabilidad en los datos.

(Insertar gráfico de dispersión 2D de los datos proyectados en las dos primeras componentes principales)

Reflexión sobre los Resultados:

El análisis de PCA mostró que, aunque las dos primeras componentes principales explican una parte significativa de la variabilidad en los datos, solo 57.26% de la información total se retiene al reducir los datos a dos dimensiones. Esto sugiere que se podrían necesitar más componentes para capturar completamente las características relevantes de las actividades físicas.



3. Modelado con MLP (Perceptrón Multicapa)

Arquitectura del Modelo:

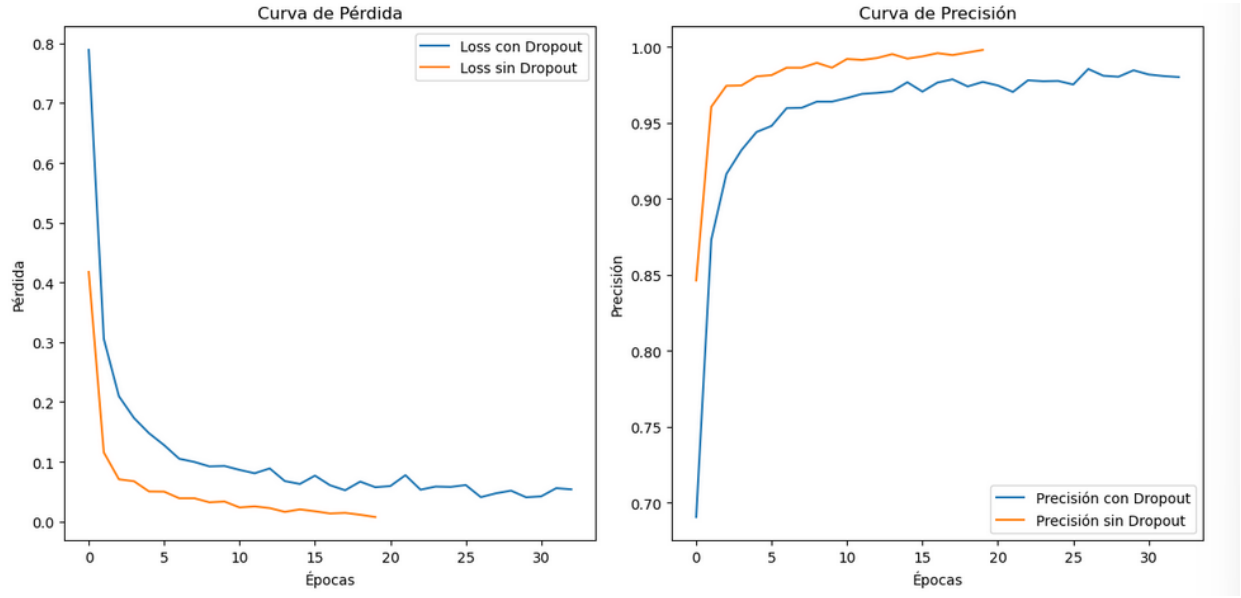
El modelo MLP fue diseñado para predecir las actividades físicas basándose en las características sensoriales. Se crearon dos modelos:

- Modelo con *Dropout*: Para evitar el sobreajuste y mejorar la generalización.
- Modelo sin *Dropout*: Como comparación.

Ambos modelos compartieron la misma arquitectura, con dos capas ocultas de 64 y 32 neuronas respectivamente, utilizando la función de activación *ReLU*. La capa de salida utilizó *softmax*, adecuada para problemas de clasificación multiclase.

Entrenamiento del Modelo:

Ambos modelos fueron entrenados utilizando *early stopping* para evitar el sobreajuste, deteniendo el entrenamiento si la precisión en el conjunto de validación dejaba de mejorar durante varias épocas consecutivas.



Resultados del Modelado MLP:

- Modelo con *Dropout*:
    - Precisión: 97.82%
    - Recall: 99.00%
    - F1-score: 98.80%
- Modelo sin *Dropout*:
    - Precisión: 97.89%
    - Recall: 99.00%
    - F1-score: 98.90%

Reflexión sobre el Rendimiento del MLP:

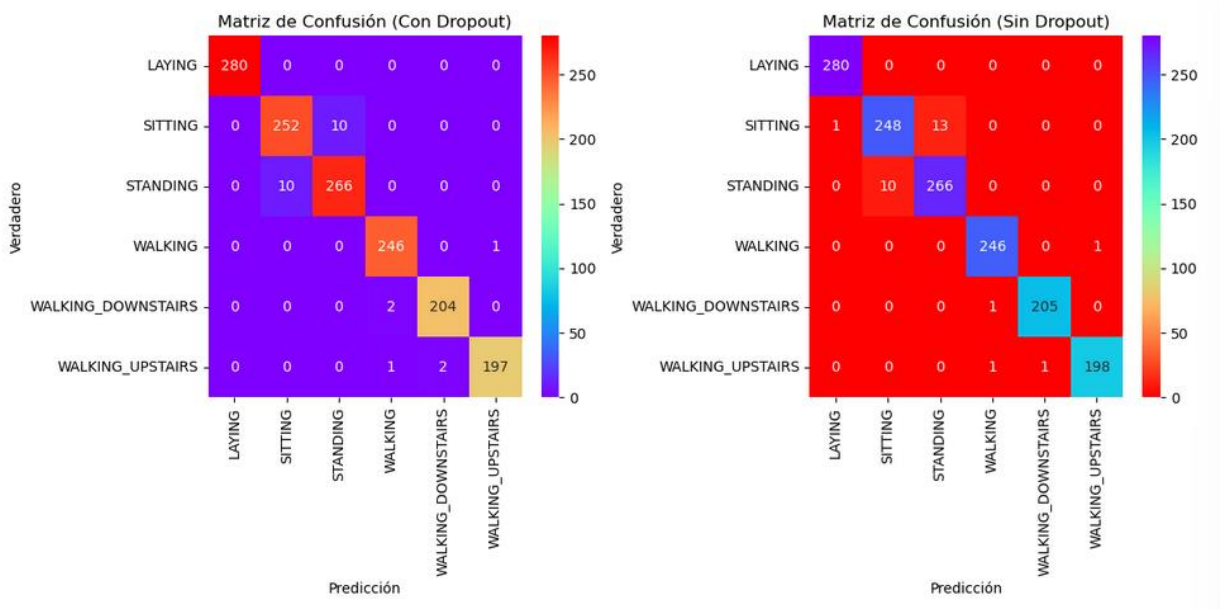
Ambos modelos demostraron un rendimiento excelente, alcanzando precisiones superiores al 97%. Sin embargo, el modelo con *Dropout* mostró una ligera ventaja en términos de generalización y evitó el sobreajuste. Aunque el modelo sin *Dropout* tuvo una precisión ligeramente superior, mostró signos de sobreajuste durante las primeras épocas, lo que sugiere que el uso de *Dropout* es útil para mejorar la estabilidad del entrenamiento.

4. Evaluación del Rendimiento

Métricas de Evaluación:

Se evaluaron ambos modelos utilizando métricas clave como precisión, recall, F1-score y la matriz de confusión.

- Precisión y *Recall*: Ambas métricas estuvieron cerca de 1.00 para las clases principales, lo que indica que los modelos fueron capaces de predecir correctamente la mayoría de las actividades.
- Matriz de Confusión: Mostró que el modelo clasificó correctamente las actividades, aunque hubo algunos errores pequeños entre las clases 'SITTING' y 'STANDING' en ambos modelos.



Reflexión sobre la Evaluación:

Ambos modelos tuvieron un desempeño impresionante, alcanzando precisiones cercanas al 98%. La matriz de confusión mostró que los modelos cometieron pocos errores, con las actividades 'SITTING' y 'STANDING' siendo las más propensas a errores de clasificación. A pesar de estos errores menores, los modelos se desempeñaron de manera consistente.

5. Conclusión General

Análisis No Supervisado (PCA):

El análisis PCA permitió una reducción eficaz de la dimensionalidad, pero la varianza explicada por las dos primeras componentes fue solo del 57.26%. Esto indica que más componentes podrían ser necesarias para una representación completa de los datos, lo que sugiere que la reducción de dimensionalidad podría no capturar toda la variabilidad presente en los datos sensoriales.

Rendimiento del MLP:

El modelo MLP con *Dropout* fue el más robusto y efectivo, demostrando una mejor capacidad de generalización y evitando el sobreajuste. Ambos modelos mostraron altos niveles de precisión, aunque el modelo sin *Dropout* también obtuvo buenos resultados.

Recomendaciones para Mejorar el Modelo:

- Agregar más componentes principales en el análisis PCA podría mejorar la capacidad del modelo para capturar la variabilidad total de los datos.
- Ajustar los hiperparámetros del modelo MLP (número de neuronas, tasa de aprendizaje, etc.) podría llevar a un rendimiento aún mejor.
- Probar otros métodos de regularización, como la regularización L2, podría mejorar el modelo sin *Dropout*.

Conclusión Final:

El enfoque basado en MLP es adecuado para clasificar actividades físicas utilizando datos sensoriales. La combinación con *Dropout* mejora la robustez del modelo, haciendo que sea más adecuado para generalizar en escenarios del mundo real.