Identificación de tipos de partículas mediante redes neuronales en datos NanoAOD del CERN

Camila Sánchez, Christopher Villota

Universidad de Nariño

Semillero de Altas Energías

camsan@udenar.edu.co, villota17@udenar.edu.co

Introducción

La identificación automática de partículas (PID) es un componente esencial en la física de altas energías, ya que permite distinguir leptones y hadrones generados en colisiones protón–protón en el LHC. El proyecto original consiste en desarrollar un clasificador basado en redes neuronales profundas utilizando directamente datos NanoAODSIM del experimento CMS en el CERN, los cuales contienen información simulada de alta fidelidad, con etiquetas de partícula provenientes del generador.

Sin embargo, antes de trabajar con los datasets completos del CERN (mucho más complejos, pesados y con estructuras internas ricas) se implementó una etapa de resultados preliminares empleando un dataset simulado más simple. Esta fase inicial permite evaluar el comportamiento general de una red neuronal multicapa (MLP) frente a la clasificación de partículas, validar el pipeline de entrenamiento, y verificar que el modelo es capaz de capturar relaciones cinemáticas fundamentales antes de escalar a los datos NanoAODSIM. De este modo, los resultados presentados aquí constituyen una primera aproximación experimental que antecede al uso del conjunto completo de NanoAODSIM del CMS.

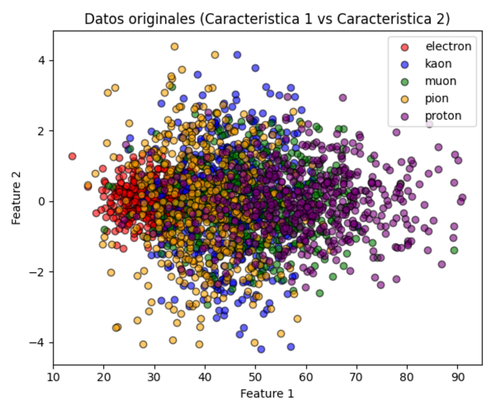
Objetivo

Evaluar el desempeño inicial de un modelo de Perceptrón Multicapa (MLP) para la identificación de partículas fundamentales. Como etapa preliminar, se empleó un *dataset* simulado con el fin de analizar la capacidad del modelo para distinguir entre distintas especies de partículas. El propósito final del proyecto es aplicar y adaptar este enfoque a datos reales del CERN en formato `NanoAOD SIM`, donde la complejidad y la variabilidad física son significativamente mayores.

Metodología

Dataset:

`particle-dataset.csv` con 6 características (el pseudo momento pT, θ, ϕ, $energia$, cal_{em} y cal_{had}) y 5 clases de partículas: electrón, muón, kaon, pion y protón.



Preprocesamiento:

- Escalamiento con `StandardScaler`.
- Codificación de etiquetas con `LabelEncoder`.
- El conjunto de datos se dividió en 80% para entrenamiento y 20% para validación.

Modelo MLP:

Se empleó una arquitectura de Perceptrón Multicapa (MLP). El diseño consideró:

- ***Arquitectura:*** Desde una red simple de dos capas ocultas hasta un modelo `FlexibleMLP` con diversos números de capas y neuronas.
- ***Funciones de activación:*** Se exploraron `ReLU`, `Sigmoid` y `Tanh` en las capas ocultas.
- ***Regularización:*** Se aplicó `Dropout` (0.3) para mitigar el sobreajuste.

Optimización y Evaluación:

El modelo fue entrenado con el optimizador `Adam` y la función de pérdida `Cross-Entropy Loss`. El desempeño se evaluó mediante *accuracy*, *precisión*, *recall*, *F1-score* y matrices de confusión.

Ajuste de Hiperparámetros:

Se realizó una búsqueda exhaustiva en cuadrícula para optimizar la *learning rate*, el *batch size*, el número de *epochs* y las dimensiones de las capas ocultas.

Búsqueda de Arquitecturas:

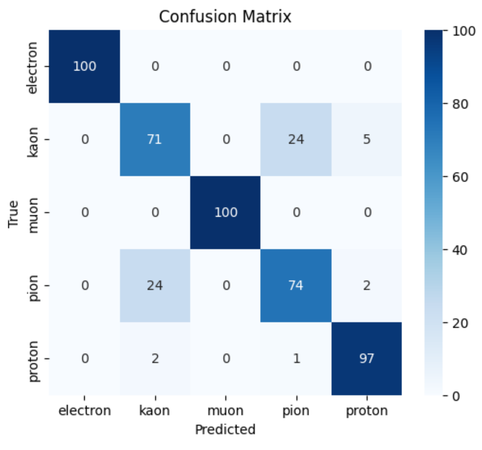
Se extendió la exploración de modelos `FlexibleMLP` con distintas dimensiones de capas y funciones de activación para encontrar una estructura óptima.

Regularización L2: Ajuste sistemático del parámetro `weight_decay`.

Resultados

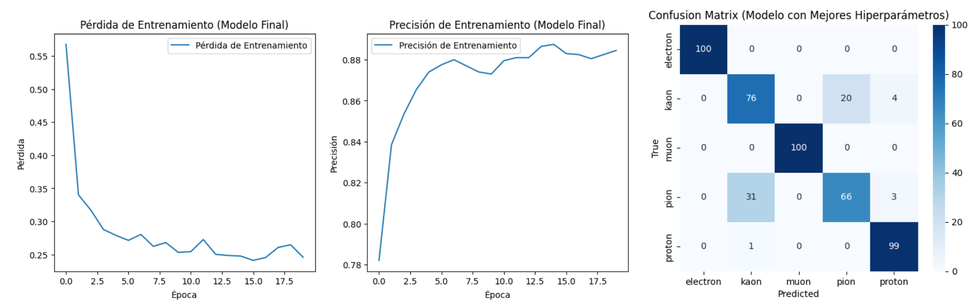
Desempeño Inicial del MLP:

El desempeño inicial del modelo MLP alcanzó una exactitud global de **0.884**, mostrando una clasificación excelente para electrones y muones (F1=**1.00**) y un rendimiento alto para protones (F1=**0.95**).
Sin embargo, las partículas *kaon* y *pion* presentaron menores puntajes F1 (0.72 y 0.74, respectivamente), lo cual se reflejó en confusiones frecuentes entre ambas clases según la matriz de confusión. Este patrón indica que sus distribuciones de características poseen un solapamiento significativo, dificultando la separación por parte del modelo.



En conjunto, estos resultados muestran que el MLP captura adecuadamente las clases más diferenciables, pero enfrenta limitaciones estructurales al distinguir partículas con propiedades más similares.

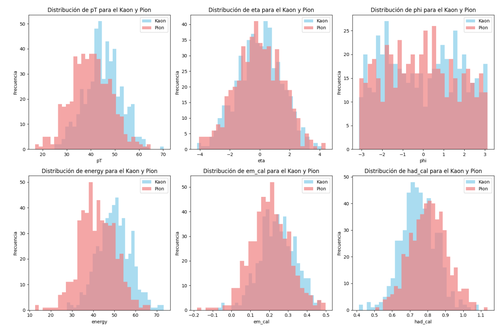
Optimización de Hiperparámetros (Grid Search): La búsqueda en cuadrícula identificó como mejores hiperparámetros `lr=0.01`, `batch_size=64`, `epochs=20`, `hidden_dim1=128` y `hidden_dim2=32`. Esta combinación produjo una exactitud de validación de **0.884**, igualando el rendimiento del modelo inicial pero proporcionando una configuración optimizada.



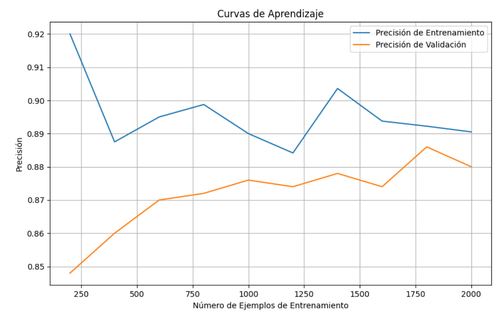
Refinamiento de arquitectura y regularización: El refinamiento adicional exploró arquitecturas diversas y regularización L2. La arquitectura óptima encontrada fue `layer_dims=[256, 128]` con función de activación `Tanh`, combinada con una regularización L2 de `weight_decay=5e-05`. Esta configuración alcanzó una exactitud de validación de **0.880**. Tras el refinamiento, los F1-scores de *kaon* y *pion* permanecieron en **0.72**, evidenciando una dificultad persistente en la clasificación de estas partículas.

Partículas Difíciles: En todas las etapas de ajuste, *kaon* y *pion* presentaron consistentemente F1-scores más bajos (aprox. 0.72–0.74) en comparación con *electron* y *muon* (ambos 1.00), y *proton* (0.96). Esto las señala como las clases más complejas para el modelo.

Análisis de Características: El menor rendimiento para estas partículas se atribuye a solapamientos importantes en sus distribuciones de características. Las estadísticas descriptivas y los histogramas de variables como pT, *energy*, *em_cal* y *had_cal* muestran que, aunque son rasgos discriminativos para otras partículas, en el caso de *kaon* y *pion* presentan distribuciones menos separables.



Curvas de Aprendizaje: La interpretación de las curvas de aprendizaje reveló una brecha persistente entre una alta exactitud en entrenamiento y una menor exactitud en validación, indicando la presencia de cierto grado de **sobreajuste**. Asimismo, se observó una meseta en las curvas, lo cual sugiere que, con la arquitectura y el conjunto de características actuales, simplemente incrementar la cantidad de datos o el número de épocas de entrenamiento no generaría mejoras significativas en el desempeño.



Conclusiones

1. El MLP clasifica correctamente las partículas electrón, muón y protón con F1-scores entre 0.96 y 1.00.
2. La distinción entre *kaon* y *pion* sigue siendo un desafío, incluso tras optimizar hiperparámetros, arquitectura y regularización.
3. Los solapamientos en características como *eta* y *phi* limitan la separabilidad entre estas clases.
4. El trabajo futuro se centrara en explorar arquitecturas de redes neuronales avanzadas (por ejemplo, aquellas específicamente diseñadas para el reconocimiento de patrones complejos), implementar técnicas de ingeniería de características más sofisticadas, o aprovechar el aprendizaje por transferencia (*transfer learning*) de modelos preentrenados en conjuntos de datos más grandes y relacionados para separar mejor las distribuciones superpuestas de las partículas 'kaón' y 'pión'.

Referencias