Predecir Clasificación de Rayos de Gama del MAGIC Gamma

María José Soto Castro

A01705840@tec.mx

INTRODUCCIÓN

Para entender cómo funcionan los telescopios como el MAGIC, es fundamental primero comprender la naturaleza de las partículas y la radiación de alta energía que están diseñados para detectar.

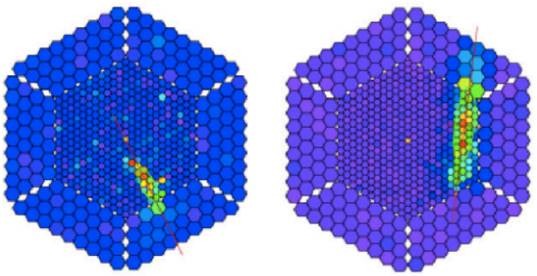
- **Rayos gamma:** Los rayos gamma son la forma más energética de radiación electromagnética, con fotones individuales que tienen energías superiores a los 100 keV. A diferencia de la luz visible, son completamente absorbidos por la atmósfera terrestre. Esto hace que su observación directa desde tierra sea imposible. Sin embargo, su interacción con la atmósfera es precisamente lo que permite a los telescopios terrestres detectarlos de forma indirecta.
- **Rayos hadrones:** Los hadrones son una clase de partículas compuestas que incluyen protones y neutrones. Los rayos hadrones, o más comúnmente, "rayos cósmicos hadrones", son hadrones de alta energía que bombardean continuamente la atmósfera terrestre. Al igual que los rayos gamma, también son absorbidos por la atmósfera, lo que inicia una cascada de partículas secundarias.

Cuando un rayo gamma de alta energía o un hadrón de rayo cósmico entra en la atmósfera, desencadena una reacción en cadena conocida como "chorro de aire extenso" (extensive air shower). Este chorro es una cascada de partículas secundarias, y es lo que los telescopios terrestres realmente observan. El desafío es que tanto los rayos gamma como los hadrones producen estos chorros, y puede ser difícil distinguirlos.

DESCRIPCIÓN DEL CONJUNTO DE DATOS

Es una colección de 19,020 instancias simuladas que representan estos chorros, cada una descrita por 10 características conocidas como parámetros de Hillas. Estos parámetros caracterizan matemáticamente la forma y orientación de la imagen del chorro, lo que permite a un algoritmo de clasificación aprender las diferencias sutiles entre los chorros regulares y elípticos de los rayos gamma y los chorros más irregulares y dispersos de los hadrones.

Una interpretación de los datos en la cámara sería algo como la siguiente imagen:



Imágenes de cámara producidas por diferentes chorros de aire extensos (EAS). Izquierda: Chorro iniciado por un rayo γ (gamma). La elipse compacta apunta a la dirección de la fuente. Derecha: Chorro hadrónico, más ancho que la elipse del chorro electromagnético y con una dirección arbitraria. (López-Oramas, Alicia., 2015)

Los datos se organizan en el dataset de la siguiente manera:

fLength: Eje mayor de la elipse.

fWidth: Eje menor de la elipse.

fSize: Logaritmo de la suma del contenido de todos los píxeles.

fConc: Relación de la suma de los dos píxeles más altos sobre el total (fsize).

fConc1: Relación del píxel más alto sobre el total (fsize).

fAsym: Distancia del píxel más alto al centro, proyectada en el eje mayor.

fM3Long: Tercera raíz del tercer momento a lo largo del eje mayor.

fM3Trans: Tercera raíz del tercer momento a lo largo del eje menor.

fAlpha: Ángulo del eje mayor con respecto al vector al origen.

fDist: Distancia del origen al centro de la elipse.

class: La clase del evento, que puede ser gamma (**g**) o hadrón (**h**).

LIMPIEZA DE DATOS

Primero, el código carga los datos brutos de un archivo llamado magic04.data y asigna nombres significativos a las columnas. A continuación, mezcla aleatoriamente todo el conjunto de datos para asegurar que la distribución de clases sea uniforme en todas las divisiones de datos. El núcleo de la limpieza implica el escalado de características: se aplica una función de normalización personalizada a todas las columnas de características (excluyendo la etiqueta de clase) para escalar sus valores a un rango entre 0 y 1. Esta normalización es crucial para algoritmos como el descenso de gradiente, que pueden funcionar mal con características en diferentes escalas. Finalmente, las etiquetas de clase, originalmente 'g' y 'h' (que representan eventos gamma y hadrón), se convierten a un formato numérico binario, con 'g' mapeado a 0 y 'h' mapeado a 1. Este preprocesamiento hace que los datos sean adecuados para el modelo de regresión logística que sigue.

MODELO UTILIZADO

El modelo de aprendizaje automático empleado para la clasificación de eventos gamma y hadrón es una implementación personalizada de regresión logística entrenada utilizando el descenso de gradiente por lotes (Batch Gradient Descent). La regresión logística es un modelo estadístico que, en su forma más básica, utiliza una función logística para modelar una variable dependiente binaria. Estima la probabilidad de que una instancia dada pertenezca a una clase particular (en este caso, un rayo gamma o un hadrón).

ENTRENAMIENTO, VALIDACIÓN Y PRUEBAS

El conjunto de datos, que contiene 19,020 instancias, se dividió en tres subconjuntos distintos para garantizar una evaluación sólida del rendimiento del modelo:

- **Conjunto de entrenamiento (60%):** Durante cada época, los parámetros del modelo se actualizaron aplicando el descenso de gradiente por lotes para minimizar la función de pérdida de entropía cruzada binaria.
- **Conjunto de validación (20%):** El error de validación se calculó en cada época para verificar el sobreajuste (overfitting) y evaluar qué tan bien el modelo se generaliza a datos no vistos. Este conjunto ayuda a determinar cuándo detener el proceso de entrenamiento para evitar que el modelo se especialice demasiado en los datos de entrenamiento.
- **Conjunto de prueba (20%):** Este conjunto no se utilizó durante el entrenamiento ni la validación y

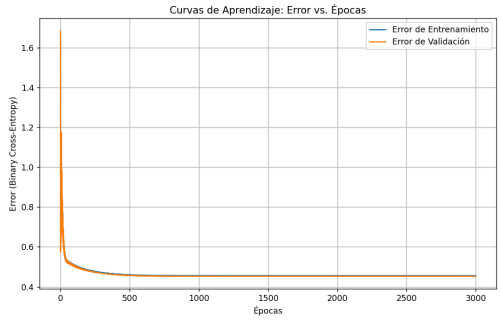
proporciona una medida imparcial del rendimiento final del modelo, específicamente su precisión en datos verdaderamente nuevos y no vistos.

El modelo fue entrenado por hasta 3,000 épocas con una tasa de aprendizaje de 0.005, y el entrenamiento se configuró para detenerse si el error de entrenamiento caía por debajo de un umbral predefinido de 0.01.

El uso de conjuntos de entrenamiento, validación y prueba separados fue un paso crítico en nuestra metodología. El conjunto de entrenamiento permitió al modelo aprender, mientras que el conjunto de validación proporcionó retroalimentación en tiempo real para ajustar el modelo y verificar el sobreajuste. La medida final e imparcial del rendimiento del modelo fue proporcionada por el conjunto de prueba, que no tuvo influencia en el proceso de entrenamiento.

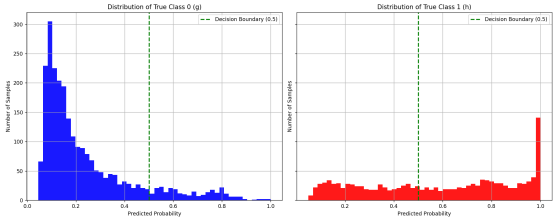
RESULTADOS

Según el análisis de las curvas de aprendizaje y los histogramas de predicción, el modelo actualmente tiene un rendimiento bajo debido a un sesgo alto. Ambas curvas de error de entrenamiento y validación convergen y se estabilizan en un valor relativamente alto de alrededor de 0.45. La distancia mínima entre estas dos curvas indica que el modelo no está memorizando los datos de entrenamiento, lo que resulta en una varianza baja.

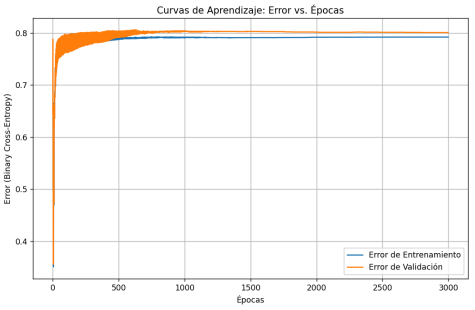


El diagnóstico de sesgo alto del modelo se apoya aún más en los histogramas de predicción. Idealmente, las distribuciones para las dos clases deberían estar bien separadas. Sin embargo, los gráficos muestran un solapamiento significativo, particularmente en el centro alrededor del límite de decisión. Esto indica que el modelo carece de la capacidad para separar las clases de manera efectiva, lo que lleva a numerosas predicciones incorrectas y a un alto error general.

También tenemos este gráfico que nos muestra cuán cerca estuvieron ambas clases de ser clasificadas como la otra. Este gráfico nos muestra que los errores más grandes se cometen al clasificar los rayos hadronicos, lo cual en este caso es favorable en lugar de los rayos gamma, ya que en este contexto es mejor confundir un rayo hadrónico con un rayo gamma.



Por último, graficamos la curva de aprendizaje en base a las épocas y podemos visualizar que el entrenamiento y la validación siguen siendo cercanas.



MEJORAS DEL MODELO Y ANÁLISIS

El diagnóstico inicial de sesgo alto en el modelo de regresión logística nos guió a través de un proceso de mejora iterativo. Identificamos que el modelo, en su forma lineal, era demasiado simple para capturar la complejidad de los datos, lo que resultó en un bajo rendimiento general.

Primera mejora: Análisis del descenso de gradiente y la regularización

Inicialmente, evaluamos la posibilidad de un problema de optimización, considerando si la tasa de aprendizaje o la regularización L2 podrían mejorar el rendimiento. Aunque la regularización L2 es una técnica efectiva para mitigar el sobreajuste, se determinó que no era la solución correcta para el problema de sesgo alto (underfitting) de nuestro modelo. La regularización, al penalizar la complejidad del modelo, habría exacerbado el problema de simplicidad, lo que confirmó la necesidad de un enfoque diferente.

También se intentó cambiar los hiper parámetros para mejorar el modelo:

Learning Rate	Precisión	Validation Accuracy
0.006	0.6149	0.7508
0.009	0.9022	0.6451
0.005	0.7425	0.8007

Se probaron varios valores para la tasa de aprendizaje, el modelo de Regresión Logística siguió demostrando un sesgo alto, lo que llevó a la necesidad de cambiar de algoritmo.

Segunda mejora: Transición a un modelo de árbol de decisión

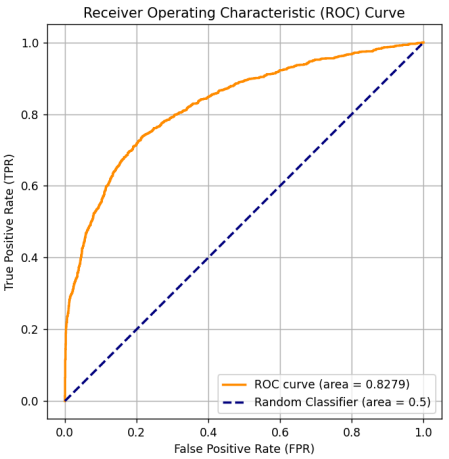
Para aumentar la capacidad del modelo de aprender relaciones no lineales, implementamos un árbol de decisión. Los resultados iniciales mostraron un sobreajuste severo:

- **Precisión de entrenamiento:** 1.00 (el modelo memorizó los datos)
- **Precisión de validación:** 0.8174

Este modelo inicial reveló inmediatamente un nuevo problema: el sobreajuste severo. Esto fue evidente por la brecha de rendimiento significativa entre la precisión de entrenamiento de 1.00 (el modelo había memorizado perfectamente los datos, lo que indica un sesgo bajo) y la precisión de validación de 0.8174 (lo que indica una varianza alta).

Para resolver el problema de sobreajuste del árbol de decisión, recurrimos a un **bosque aleatorio (Random Forest)**, que utiliza el concepto de *bagging*. El bosque aleatorio entrena múltiples árboles de decisión pequeños en subconjuntos de datos aleatorios y promedia sus predicciones. También realizamos ajustes colocando 250 árboles en el conjunto para evitar el sobreajuste.

RESULTADOS FINALES



Análisis de la Curva ROC y el AUC

La Curva Característica Operativa del Receptor (ROC) y el Área bajo la Curva (AUC) son herramientas fundamentales para evaluar el rendimiento de un modelo de clasificación, proporcionando un análisis visual y cuantitativo de su capacidad para distinguir entre clases.

Esta gráfica representa la relación entre la Tasa de Verdaderos Positivos (TPR) y la Tasa de Falsos Positivos (FPR) en diversos umbrales de decisión. Una curva que se inclina marcadamente hacia la esquina superior izquierda indica que el modelo es capaz de identificar correctamente la mayoría de los casos positivos (alta TPR) sin cometer un gran número de errores (baja FPR).

El Área bajo la Curva (AUC): El valor de AUC resume el rendimiento del modelo en una sola métrica. Un AUC de 1.0 representa un modelo perfecto, mientras que un AUC de 0.5 corresponde a un clasificador aleatorio.

Interpretación de los Resultados (AUC = 0.8279):

El valor de AUC de 0.8279 indica que el modelo posee un buen poder discriminatorio. La curva de el modelo se encuentra muy por encima de la línea del clasificador aleatorio, lo que confirma que es significativamente mejor que el azar para distinguir entre los rayos gamma y los hadrones. Este resultado valida la efectividad del modelo de Random Forest y complementa los altos valores de precisión obtenidos en las pruebas.

CONCLUSIÓN

El análisis y la mejora de este modelo de clasificación de rayos gamma y hadrones demuestran la importancia de un enfoque metódico en el aprendizaje automático. El proyecto comenzó con un modelo de Regresión Logística, que si bien es simple y robusto, fue diagnosticado con sesgo alto (underfitting), lo que limitaba su capacidad para capturar las relaciones complejas de los datos.

A través de un proceso iterativo, pasamos a un Árbol de Decisión, que soluciona el problema de sesgo pero introdujo varianza alta (overfitting). Finalmente, el uso de un Random Forest demostró ser la solución más efectiva. Al combinar múltiples árboles en un método de conjunto, el modelo logró mitigar la alta varianza y generalizar de manera superior a datos no vistos.

Los resultados del modelo final, con una precisión y un AUC significativamente altos, validan la elección de esta arquitectura. En última instancia, este proyecto subraya que el éxito en la construcción de modelos no reside en la elección de un solo algoritmo, sino en la capacidad de diagnosticar con precisión el rendimiento del modelo y aplicar las técnicas de mejora adecuadas para lograr el mejor ajuste y la mayor capacidad de generalización.

REFERENCIAS

Conceição, R., Costa, P. J., Gibilisco, L., Pimenta, M., & Tomé, B. (2023). *: a high resolution gamma/hadron and composition discriminant variable for Water-Cherenkov Detector cosmic-ray observatories*

MAGIC – INAF Osservatorio Astronomico di Brera. (2019). Inaf.it. https://brera.inaf.it/en/progetti_ricerca/magic/#:~:text=SHORT%20DESCRIPTION,physics%20and%20high%2Denergy%20astrophysics.

Probing the Cosmic Ray Background of Gamma-Ray Astronomy with Hadron Colliders. (2017). Arxiv.org. https://arxiv.org/html/2509.04040v1#:~:text=These%20CR%20particles%2C%20composed%20predominantly,form%20of%20hadronic%20air%20showers.

López-Oramas, Alicia. (2015). *Multi-year Campaign of the Gamma-Ray Binary LS I +61◦ 303 and Search for VHE Emission from Gamma-Ray Binary Candidates with the MAGIC Telescopes*. 10.13140/RG.2.1.4140.4969.