



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Dirigido a la Comisión de Servicio Social de la Licenciatura en Física

Informe Final de Actividades de Servicio Social

Aplicación de Métodos de Machine Learning para la Clasificación de Eventos en Colisiones de Partículas usando Datos Abiertos del CERN

Nombre: Ahuatzi Pichardo Mariano Josué

Nombre del Tutor: Leonid Serkin

Número de Cuenta: 313145803

Carrera: Física

Clave del Programa: SS-2025-12 / 139-444

Proyecto: Apoyo a la Investigación

Periodo de Servicio Social: 25 / 04 / 2025 - 01 / 02 / 2026

Facultad de Ciencias, UNAM, Ciudad Universitaria,
Círculo Exterior S/N, C.P. 04510
Febrero del 2026

Índice

1. Introducción: Comprensión del problema	2
2. Descripción de los datos	3
2.1. Detector ATLAS	3
2.2. El bosón de Higgs	4
3. Desarrollo	6
3.1. Entorno de desarrollo	6
3.2. Preprocesamiento de datos	6
3.3. Histogramas	8
3.4. Modelado base	8
3.5. Esquema de validación, selección de Características y entrenamiento	8
3.6. Optimización de hiperparámetros	9
3.7. Evaluación	10
4. Resultados	11
4.1. Métricas del modelo optimizado	13
5. Resultados para la carrera de física	16
6. Discusión y Conclusiones	16
A. Diccionario de datos	18

1. Introducción: Comprensión del problema

En la Física de Altas Energías Contemporánea, el estudio del bosón de Higgs constituye uno de los principales retos de la investigación actual. El Modelo Estándar, la teoría más exitosa hasta la fecha para describir las interacciones entre las partículas elementales, ha demostrado una notable capacidad predictiva. Sin embargo deja abiertas cuestiones fundamentales, entre ellas la determinación de la masa del bosón de Higgs y la caracterización completa de sus propiedades. Estas limitaciones hacen que el análisis detallado del bosón siga siendo un tema de gran interés tanto teórico como experimental.

En 2012, los experimentos ATLAS y CMS del CERN confirmaron por primera vez la existencia del bosón de Higgs, al alcanzar una significancia estadística de 5σ , equivalente a un nivel de confianza cercana a 0.9999994 en el análisis estadístico que probaba su existencia. Este descubrimiento, se obtuvo mediante técnicas de reconstrucción de eventos y pruebas de hipótesis aplicadas a datos de colisiones protón–protón a energías de 8 TeV . Este suceso, en su momento, representó un hito histórico para la física de partículas.

Como se mencionó, el Modelo Estándar no proporciona predicciones completas sobre todas las propiedades del bosón de Higgs, por lo que su estudio continúa siendo un área activa y esencial dentro de la investigación actual.

El principal desafío radica en que la señal asociada al bosón de Higgs tras su desintegración se encuentra inmersa en un entorno altamente ruidoso: cada colisión genera miles de partículas y procesos secundarios, y sólo una fracción minúscula corresponde realmente a eventos vinculados al bosón de Higgs. Distinguir estas señales es comparable a encontrar una “aguja” dentro de un vasto “pajar” de fondo, lo cual constituye un problema estadístico complejo que exige técnicas avanzadas de discriminación.

En este contexto, y aprovechando que los experimentos actuales operan a energías mayores (13 TeV), generan grandes volúmenes de datos más completos y se producen simulaciones de estos eventos, surge la oportunidad de emplear métodos modernos de aprendizaje automático para mejorar la separación entre señal y fondo. El uso de modelos de clasificación basados en *machine learning* permite capturar relaciones complejas entre variables cinemáticas, ofreciendo potencialmente una capacidad superior para identificar eventos compatibles con la producción del bosón de Higgs a un menor coste computacional.

Este proyecto tiene como objetivo desarrollar un modelo de *Machine Learning* capaz de distinguir de manera eficaz la señal asociada al bosón de Higgs del ruido de fondo. Asimismo, se implementan estrategias orientadas a la optimización de métricas de clasificación relevantes, con el fin de mejorar la clasificación de los eventos vinculados a la producción del bosón de Higgs. De esta forma, contar con una herramienta de clasificación robusta permite aislar de manera más precisa los eventos de interés y, en consecuencia, contribuye a una mejor comprensión de sus propiedades físicas.

2. Descripción de los datos

Los datos utilizados en este proyecto provienen del conjunto de Datos Abiertos del CERN de 13 TeV [5]. Este material se pone a disposición del público con fines educativos y de investigación, y consiste en simulaciones de colisiones protón–protón reconstruidas por el detector ATLAS. Cada registro, denominado evento, contiene la información completa asociada a una colisión individual, incluyendo las partículas producidas, sus trayectorias y las mediciones realizadas por los distintos subsistemas del detector.

2.1. Detector ATLAS

El experimento ATLAS es un detector de propósito general instalado en el Gran Colisionador de Hadrones (LHC) en el CERN. Su función es registrar y analizar las partículas producidas en colisiones protón–protón para estudiar fenómenos del Modelo Estándar y buscar nueva física más allá del Modelo Estandar

ATLAS está construido en capas concéntricas alrededor del punto de interacción, de modo que cada subsistema cumple una función específica en la detección [1] como se muestra en la Figura 1. El Detector Interno registra con alta precisión las trayectorias de partículas cargadas cerca del punto de colisión; los calorímetros electromagnético y hadrónico miden la energía depositada por electrones, fotones y hadrones; y el espectrómetro de muones, situado en la periferia, detecta muones que atraviesan prácticamente todas las capas previas. Todo el conjunto está inmerso en potentes campos magnéticos que permiten curvar las trayectorias de partículas cargadas para determinar su momento. Las trayectorias de las partículas así como el proceso de detección a través de los subsistemas del detector se muestran en un corte transversal del mismo en la Figura 2

La combinación de todos estos subsistemas hace que ATLAS sea esencialmente un detector “hermético”, capaz de reconstruir de manera casi completa los eventos generados en las colisiones. Esto permite identificar partículas, medir energías y momentos, reconstruir vértices de decaimiento y analizar firmas experimentales complejas, como las asociadas a la producción del bosón de Higgs u otros procesos. Gracias a esta arquitectura, ATLAS puede estudiar desde interacciones conocidas hasta señales potencialmente indicativas de nueva física.

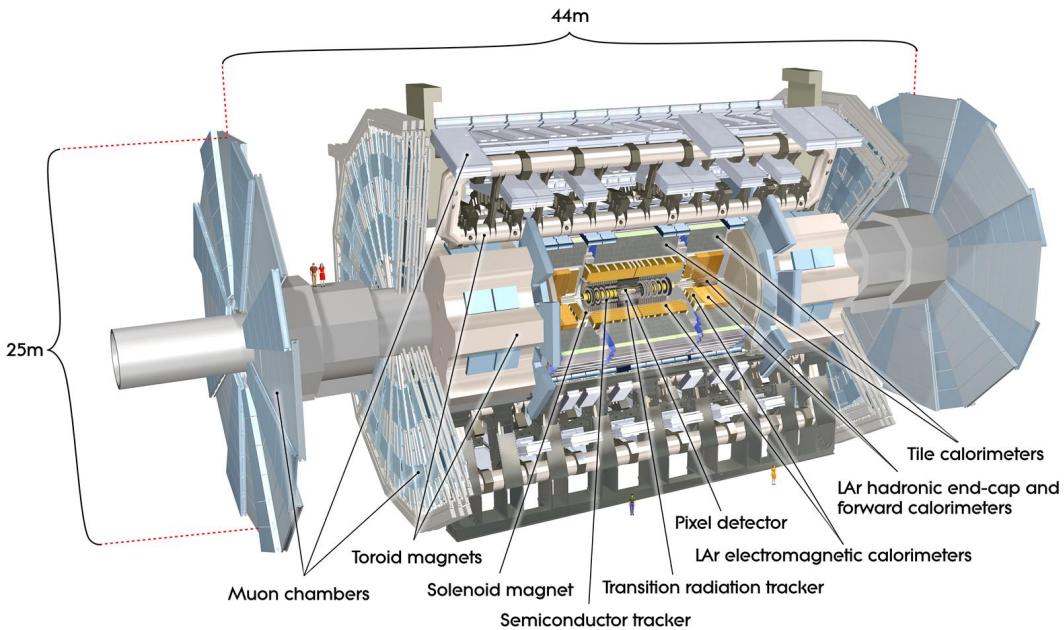


Figura 1: Detectores (subsistemas) del Experimento Atlas

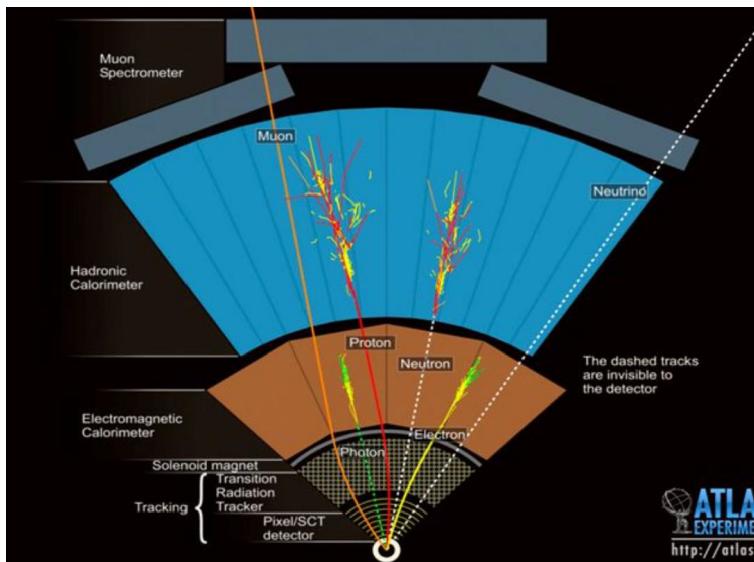


Figura 2: Corte Transversal a los Detectores de Atlas y Detección de Partículas

2.2. El bosón de Higgs

En las colisiones protón–protón del LHC se producen una gran variedad de partículas. Entre ellas el bosón de Higgs, cuya producción dominante ocurre mediante fusión de gluones, seguida del decaimiento en diversos canales [4]. Para un Higgs con masa $m_H = 125 \text{ GeV}$, los principales modos de decaimiento presentan fracciones de ramificación bien conocidas: $H \rightarrow b\bar{b}$ (58 %), $H \rightarrow WW^*$ (21 %), $H \rightarrow gg$ (8.6 %), $H \rightarrow \tau^+\tau^-$ (6.3 %),

$H \rightarrow ZZ^*$ (2.6 %), $H \rightarrow \gamma\gamma$ (0.23 %), entre otros [8]. Sin embargo sólo los canales leptónicos y fotónicos (aunque menos frecuentes), permiten señales más limpias en el detector lo que en los datos se traduce como más fáciles de analizar.

En este proyecto se analiza específicamente el canal

$$H \rightarrow WW^* \rightarrow \ell\nu\ell\nu,$$

en el cual cada bosón W decae en un lepton cargado (electrón o muón) y un neutrino. Este canal presenta una tasa de decaimiento moderada y una firma experimental bien definida, caracterizada por la presencia de dos leptones cargados y una cantidad de energía transversal faltante asociada a los neutrinos no detectados.

No obstante, este no es el único proceso en colisiones protón–protón que puede dar lugar a un estado final con dos leptones y energía faltante. Entre los principales procesos de fondo se encuentran la producción no resonante de pares de bosones WW , los decaimientos del bosón Z a leptones tau ($Z \rightarrow \tau\tau$), la producción de pares de quarks top ($t\bar{t}$), así como procesos $W+jets$ con leptones no genuinos. Todos estos procesos pueden producir estados finales experimentalmente indistinguibles del canal de señal, lo que motiva el uso de técnicas avanzadas de selección y clasificación para mejorar la separación entre señal y fondo.

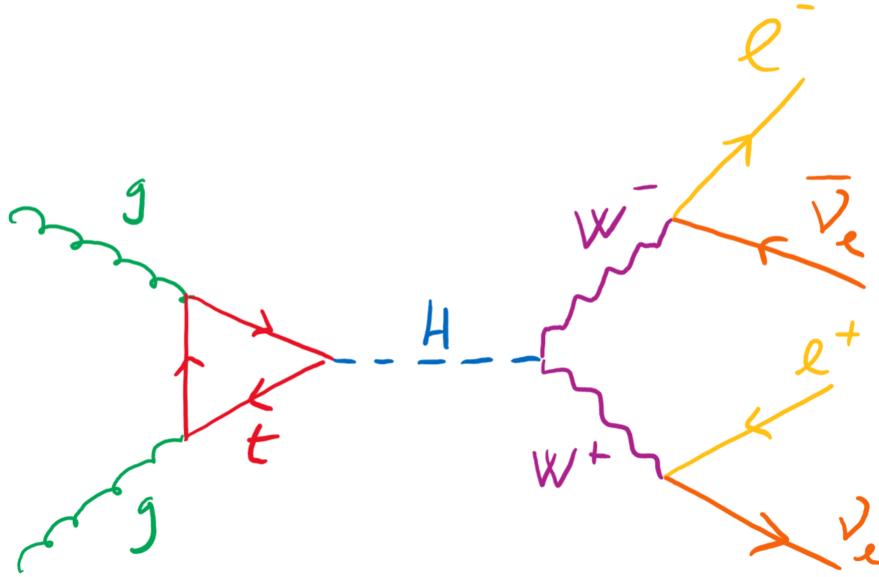


Figura 3: Diagrama de Feynmann del proceso ggF : $gg \rightarrow H \rightarrow WW^* \rightarrow \ell\ell\nu\nu$, donde un Higgs producido por fusión de gluones decae en dos leptones y neutrinos.

3. Desarrollo

3.1. Entorno de desarrollo

Para el desarrollo de este proyecto, configuramos el entorno de trabajo siguiendo el esquema híbrido propuesto por ATLAS Open Data [6]. Se instaló el contenedor de Docker [7] recomendado en dicha referencia, desplegándolo en un entorno local para encapsular las dependencias, bibliotecas y configuraciones necesarias para el análisis.

Dentro del contenedor desplegamos un entorno basado en Python para el procesamiento y análisis de datos, integrando herramientas del ecosistema ROOT (framework desarrollado por el CERN para en análisis de física de altas energias) [10] para la lectura de archivos en formato .root. La configuración incluyó la instalación de las dependencias científicas necesarias y bibliotecas especializadas en *Machine Learning* principalmente *scikit-learn*.

3.2. Preprocesamiento de datos

Descargamos los archivos correspondientes a las señales discutidas en la sección 2.2 disponibles en [5], estos datos corresponden a señales con exactamente 2 leptones detectados, con las siguientes etiquetas y nombres de archivos:

- Higgs: mc_345324.ggh125_WW2lep.exactly2lep.root
- Dibosón: mc_363492.llvv.exactly2lep.root
- ttbar: mc_410011.single_top_tchan.exactly2lep.root
- Ztautau: mc_361108.Ztautau.exactly2lep.root

Implementamos una función encargada de procesar los archivos en formato ROOT y transformar los datos a una estructura tabular. En esta etapa realizamos un primer filtrado, conservando únicamente eventos con exactamente dos leptones y como máximo un jet, una seleccionar únicamente las variables relevantes para los cortes físicos recomendados y aquellas variables cinemáticas que posteriormente se utilizarán en el entrenamiento del modelo, además del cálculo de cantidades cinemáticas derivadas: la masa invariante dileptónica (m_{LL}), el momento transversal del sistema dileptónico (p_T^L), y diferencias angulares ($\Delta\phi_{ll}$, $\Delta\phi_{ll,MET}$). Las variables seleccionada se discuten a profundidad en la sección: A.

Las variables con múltiples entradas por evento fueron desagregadas en el número de columnas necesarias para garantizar su correcta representación en formato tabular. Con estas transformaciones, los datos fueron almacenados en un DataFrame de Pandas, para su manipulación y análisis posterior.

Una vez completada esta etapa inicial, aplicamos los cortes físicos establecidos en la tabla 3.5 de [3], con el objetivo de seleccionar eventos candidatos a provenir de la desintegración del bosón de Higgs. Este procedimiento se realizó para todas las muestras consideradas (señal y fondos). Estos cortes fueron:

1. **Disparador (trigger):** evento activado para **un electrón o muón aislado.**
2. **Leptones: Exactamente Dos Leptones** aislados (electrones o muones) de distinto sabor y signo opuesto:
 - a) $p_T(e) > 22 \text{ GeV}$, $p_T(\mu) > 15 \text{ GeV}$.
 - b) Aislamiento más estricto: $\text{lep etcone20} < 0.1$, $\text{lep ptcone30} < 0.1$.
 - c) indica si el leptón satisface los criterios estrictos de identificación $\text{lep isTightID} = \text{True}$
3. **Energía transversal faltante:**
 - a) $E_T^{\text{miss}} > 30 \text{ GeV}$ (indicando la presencia de neutrinos)
4. **Jets:**
 - a) Cero o a lo mas un jet (i.e 0 o 1) con $p_T > 30 \text{ GeV}$.
 - b) El evento no debe contener ningún jet etiquetado como proveniente de un quark b (b-tagged jet) con $p_T > 20 \text{ GeV}$
 - c) (lo anterior, otra vez) Ningún jet etiquetado como *b-jet* (veto de *b-jets*) para suprimir el fondo de $t\bar{t}$.
5. **Ángulos azimutales:**
 - a) Entre el sistema dileptónico y el momento transversal faltante: $\Delta\phi(\ell\ell, E_T^{\text{miss}}) > \pi/2$.
 - b) Entre los dos leptones: $\Delta\phi(\ell, \ell) < 1.8$.
6. **Impulso transversal del sistema dileptónico:** $p_T^{\ell\ell} > 30 \text{ GeV}$.
7. **Masa invariante del par leptónico:** $10 \text{ GeV} < m_{\ell\ell} < 55 \text{ GeV}$.

Posteriormente, graficamos la comparación entre el número total de eventos originales y aquellos que superaron el preprocesamiento. Finalmente, los eventos seleccionados se guardaron en un archivo único llamado `Datos Unidos.csv` para su uso en las etapas de modelado y análisis estadístico.

3.3. Histogramas

De acuerdo con los resultados obtenidos posteriores a la aplicación de filtros, seleccionamos únicamente las dos señales que cuentan con un número suficiente de eventos para realizar un análisis estadísticamente significativo y entrenar un modelo de clasificación. Estas corresponden a la señal asociada a la desintegración del bosón de Higgs y fondo dominante proveniente de la producción de dibosones

Se implementó una función para visualizar la distribución de variables físicas mediante histogramas normalizados que comparan señal y fondo. La función separa automáticamente los eventos según su etiqueta de clase y utiliza intervalos (bins) comunes para garantizar una comparación consistente un histograma representativo se muestra en la sección de 4

La normalización permite analizar la forma de las distribuciones independientemente del número total de eventos, facilitando la evaluación del poder discriminante de cada variable. Además, la función automatiza el guardado de las figuras.

3.4. Modelado base

Implementamos un análisis preliminar de Importancia de Variables basado en el clasificador `RandomForestClassifier` ordenandolas de forma descendiente de acuerdo a su nivel de significancia en la clasificación entre señal y fondo.

Implementamos un entrenamiento inicial entre modelos de clasificación representativos de distintas familias. Aplicamos métodos de bagging (Random Forest), boosting clásico (Gradient Boosting), así como enfoques de gradient boosting optimizado (LightGBM y XGBoost). Como modelos no basados en árboles, incluimos una SVM con kernel RBF y una regresión logística como modelo lineal. En los modelos que lo requieren, incorporamos un escalamiento estándar mediante un *pipeline* para asegurar comparabilidad y estabilidad numérica.

La evaluación se realizó mediante validación cruzada estratificada de 5 particiones (`StratifiedKFold`), preservando la proporción señal/fondo en cada *fold*. Para cada modelo calculamos métricas complementarias: *accuracy* (desempeño global), *AUC-ROC* (capacidad de separación independiente del umbral) y *F1-score* (balance entre precisión y exhaustividad en la clase positiva). Finalmente, reportamos la media y desviación estándar de cada métrica a través de los *folds*, con el fin de cuantificar tanto el rendimiento como su estabilidad.

3.5. Esquema de validación, selección de Características y entrenamiento

Para obtener una estimación más robusta y reproducible del desempeño de los modelos, implementamos un esquema de validación cruzada estratificada basado en *folds*

persistidos en archivos. Para ello, a partir del conjunto unificado (`Datos Unidos.csv`) generamos 5 particiones mediante `StratifiedKFold`, conservando la proporción señal/-fondo en cada división. En cada iteración se construyen explícitamente los conjuntos de entrenamiento y validación correspondientes y se guardan como archivos CSV independientes (`fold_i_train.csv` y `fold_i_val.csv`). Esta estrategia permite reutilizar exactamente las mismas particiones a lo largo de todo el flujo experimental (comparación de modelos, selección de variables y optimización).

Adicionalmente, para la selección de características adoptamos un enfoque basado en validación cruzada: en cada *fold* entrenamos un clasificador LightGBM sobre el conjunto de entrenamiento y extraemos la importancia de cada variable. Posteriormente, promediamos dichas importancias entre los cinco *folds* para obtener un ranking global estable, reduciendo la dependencia respecto a una única partición. Con este criterio seleccionamos las k variables más relevantes ($k=16$), priorizando observables con contribución consistente a lo largo de toda la validación cruzada. Este procedimiento fortalece el entrenamiento al disminuir el riesgo de sobreajuste por selección de variables y al mantener únicamente atributos con poder discriminante sostenido entre señal y fondo.

Entrenamos nuevamente los modelos previamente descritos utilizando únicamente las variables seleccionadas como más significativas. Posteriormente, evaluamos su desempeño mediante el mismo esquema de validación cruzada estratificada, obteniendo las métricas de evaluación más relevantes. Los resultados comparativos se presentan en la Sección 4.

3.6. Optimización de hiperparámetros

Considerando el desempeño observado de los modelos anteriores se decidió optimizar el modelo LightGBM, para esto implementamos un proceso de optimización de hiperparámetros utilizando Optimización Bayesiana a través de la biblioteca Optuna [9]. A diferencia de métodos exhaustivos como Grid Search, la optimización bayesiana explora de manera inteligente el espacio de búsqueda, modelando la función objetivo y priorizando combinaciones de hiperparámetros con mayor probabilidad de mejora [11].

Para la evaluación del desempeño del modelo se definió una función objetivo que, para cada conjunto de hiperparámetros propuesto por el algoritmo, entrena un modelo LightGBM y evalúa su desempeño mediante validación cruzada estratificada de 5 particiones, utilizando como métrica objetivo el área bajo la curva ROC (AUC). El espacio de búsqueda incluyó parámetros estructurales del modelo (como `num_leaves`, `max_depth` y `min_data_in_leaf`), así como parámetros de regularización y muestreo (`feature_fraction`, `bagging_fraction`, `lambda_l1`, `lambda_l2`). El estudio se ejecutó durante 40 iteraciones, maximizando la métrica ROC-AUC promedio en validación cruzada.

Una vez identificados los mejores hiperparámetros, se entrenó el modelo optimizado manteniendo constantes otros parámetros globales (como el número de estimadores y la tasa de aprendizaje), asegurando comparabilidad con el modelo base.

3.7. Evaluación

Una vez seleccionadas las variables más relevantes y optimizados los hiperparámetros del modelo LightGBM, realizamos la evaluación final utilizando exclusivamente el subconjunto reducido de características. Para garantizar una estimación robusta y comparable con etapas previas, empleamos nuevamente un esquema de validación cruzada estratificada de cinco particiones, preservando la proporción señal/fondo en cada fold.

Se calcularon las métricas de desempeño más relevantes para clasificación binaria: Accuracy, AUC-ROC y F1-score, en el mismo formato de las evaluaciones previas, reportando la media y desviación estándar entre particiones. Este procedimiento permite cuantificar no solo el rendimiento promedio del modelo, sino también su estabilidad frente a diferentes divisiones del conjunto de datos.

Adicionalmente, se generaron predicciones probabilísticas fuera de muestra (out-of-fold) mediante validación cruzada, lo que permitió evaluar el comportamiento del modelo sin introducir sesgo de sobreajuste. A partir de estas probabilidades se construyó la curva ROC, obteniendo el área bajo la curva (AUC) como medida global de separación entre señal y fondo. De forma complementaria, se calculó la curva Precision–Recall y el Average Precision (AP), métricas particularmente informativas en escenarios con desbalance moderado de clases.

También se obtuvo la matriz de confusión normalizada, permitiendo evaluar las tasas de clasificación correcta por clase y el equilibrio entre falsos positivos y falsos negativos. Finalmente, se examinó la distribución del score probabilístico del modelo para señal y fondo, así como la curva de calibración, con el fin de verificar la calidad probabilística de las predicciones y su alineación con una calibración ideal.

4. Resultados

La cantidad de datos obtenidos después del procesamiento se muestra en la siguiente gráfica, muestra un ligero desbalance de clases entre la señal proveniente del bosón de Higgs y la señal proveniente del dibosón WW.

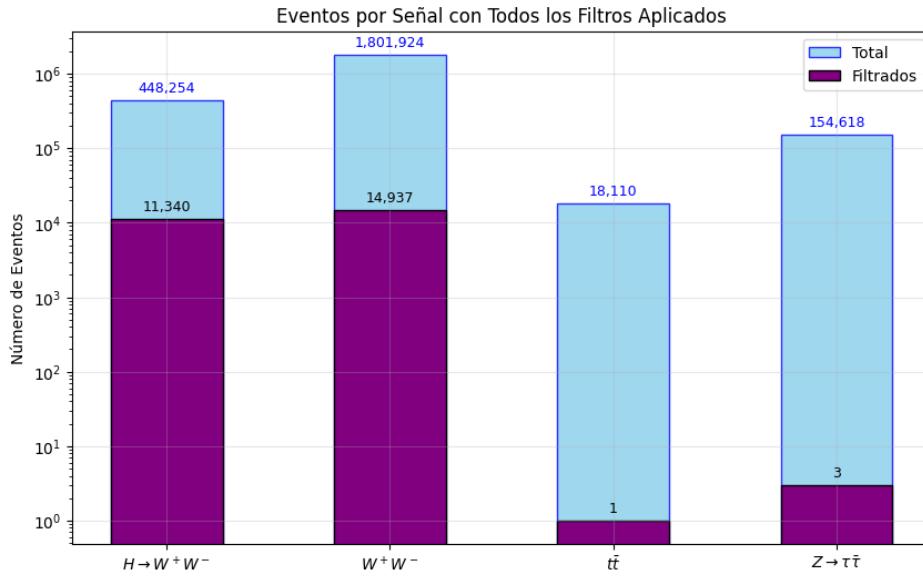


Figura 4: Cantidad de eventos antes y después de aplicar preprocesamiento y filtrado

Histogramas Normalizado de la masa invariante del sistema dileptónico para los procesos correspondientes a la señal del bosón de Higgs y al fondo dominante de dibosones

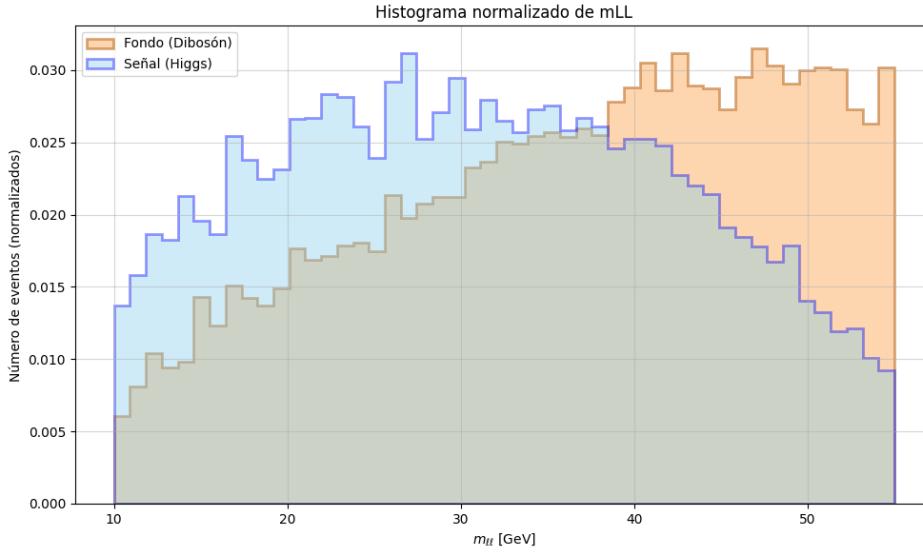


Figura 5: Histograma normalizado de masa la masa invariante del sistema dileptónico para los procesos provenientes de la señal Higgs y fondo dibosón

El análisis de selección de características muestra que la energía transversal faltante es la variable con mayor capacidad discriminante entre señal y fondo, seguida por la masa invariante dileptónica y el momento transversal del sistema dileptónico, que capturan información clave sobre la cinemática y la topología del decaimiento del bosón de Higgs. El conjunto de datos presenta un desbalance moderado (43 % señal y 57 % fondo), lo que justifica el uso de validación cruzada estratificada.

Métricas de desempeño el entrenamiento de los modelos base con una validación cruzada de 5 pliegues

Cuadro 1: Desempeño de los modelos base mediante validación cruzada estratificada (media ± desviación estándar).

Modelo	Accuracy	AUC	F1-score
Gradient Boosting	0.679 ± 0.003	0.747 ± 0.004	0.633 ± 0.007
LightGBM	0.677 ± 0.004	0.743 ± 0.005	0.633 ± 0.007
Random Forest	0.676 ± 0.003	0.743 ± 0.003	0.617 ± 0.004
XGBoost	0.673 ± 0.006	0.741 ± 0.004	0.627 ± 0.007
SVM (RBF)	0.678 ± 0.004	0.740 ± 0.004	0.636 ± 0.005
Logistic Regression	0.666 ± 0.004	0.732 ± 0.008	0.598 ± 0.006

Cuadro 2: Ranking de importancia (número promedio de veces que la variable fue utilizada para dividir un nodo en un arbol) promedio de variables obtenido mediante validación cruzada con LightGBM.

Variable	Importancia Media
met_et	757.0
dphi_ll_met	611.8
mLL	610.6
pTll	505.4
lep_etcone20_0	500.6
lep_pt_1	475.6
lep_pt_0	466.8
lep_etcone20_1	464.2
lep_E_0	420.6
dphi_ll	412.4
lep_E_1	405.4
met_phi	402.8
jet_pt	381.8
lep_phi_1	375.6
lep_eta_0	369.2
lep_phi_0	364.0

Cuadro 3: Comparación de desempeño de los modelos entrenados con selección de características mediante validación cruzada estratificada (media ± desviación estándar).

Modelo	Accuracy	AUC	F1-score
Gradient Boosting	0.679 ± 0.003	0.747 ± 0.004	0.634 ± 0.007
LightGBM	0.678 ± 0.005	0.744 ± 0.004	0.636 ± 0.006
SVM (RBF)	0.679 ± 0.007	0.742 ± 0.004	0.640 ± 0.007
XGBoost	0.674 ± 0.004	0.741 ± 0.004	0.628 ± 0.007
Random Forest	0.674 ± 0.003	0.741 ± 0.005	0.623 ± 0.004
Logistic Regression	0.666 ± 0.002	0.731 ± 0.008	0.598 ± 0.005

4.1. Métricas del modelo optimizado

Cuadro 4: Desempeño Final del modelo LightGBM Optimizado con conjunto de Características Seleccionadas (media ± desviación estándar en validación cruzada estratificada).

Métrica	Valor
Accuracy	0.680 ± 0.003
AUC-ROC	0.749 ± 0.005
F1-score	0.636 ± 0.007

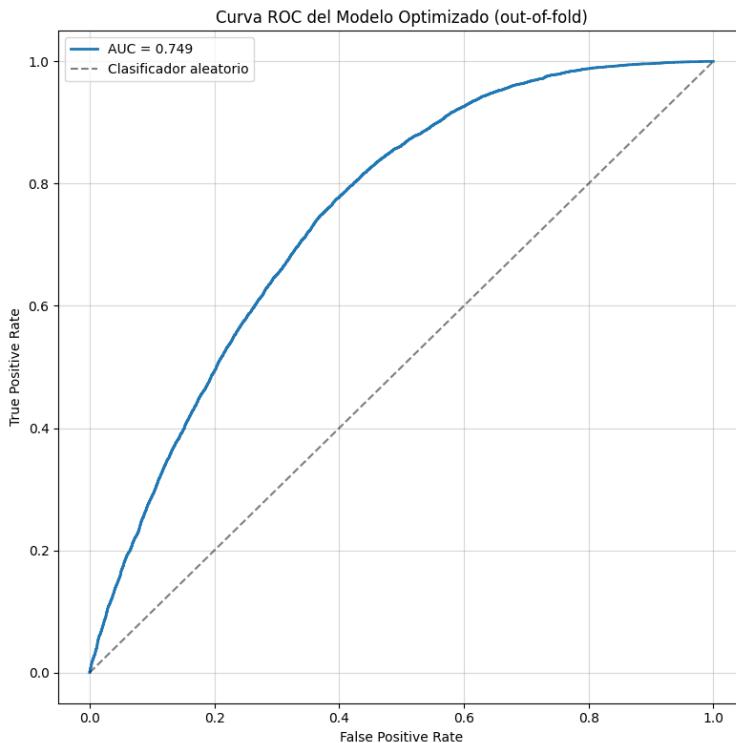


Figura 6: Curva ROC

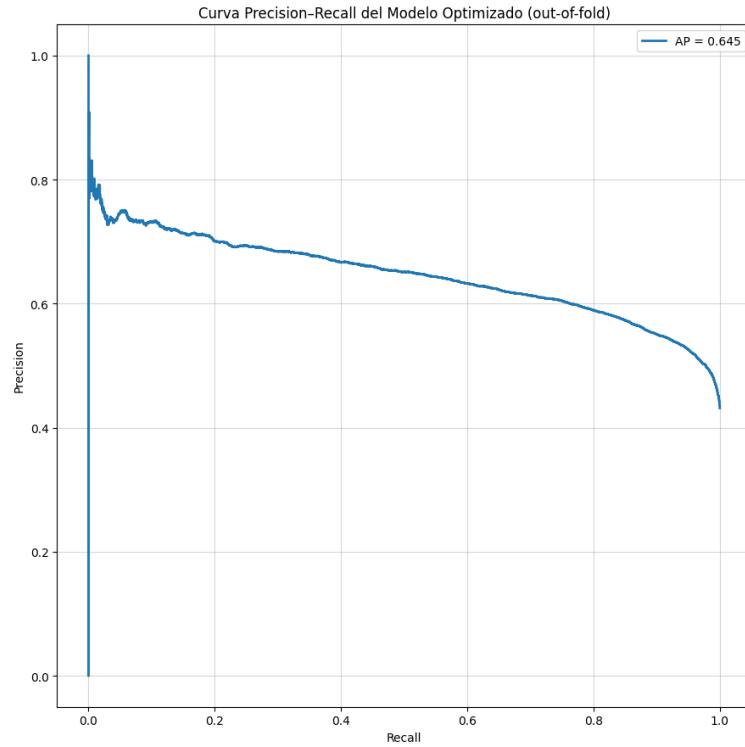


Figura 7: Precision-Reall

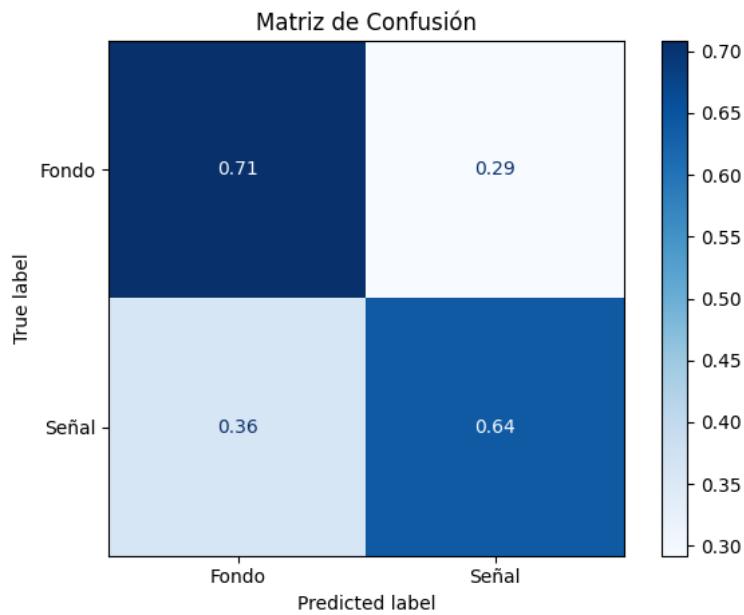


Figura 8: Matriz de confusión

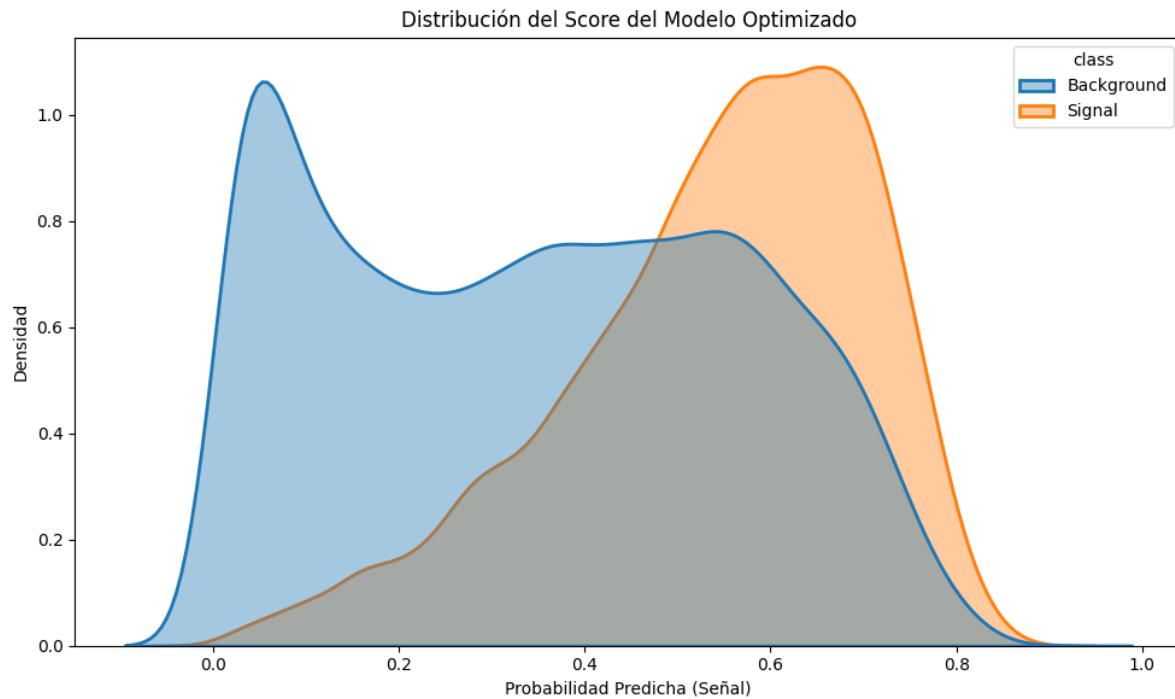


Figura 9: Distribución del score

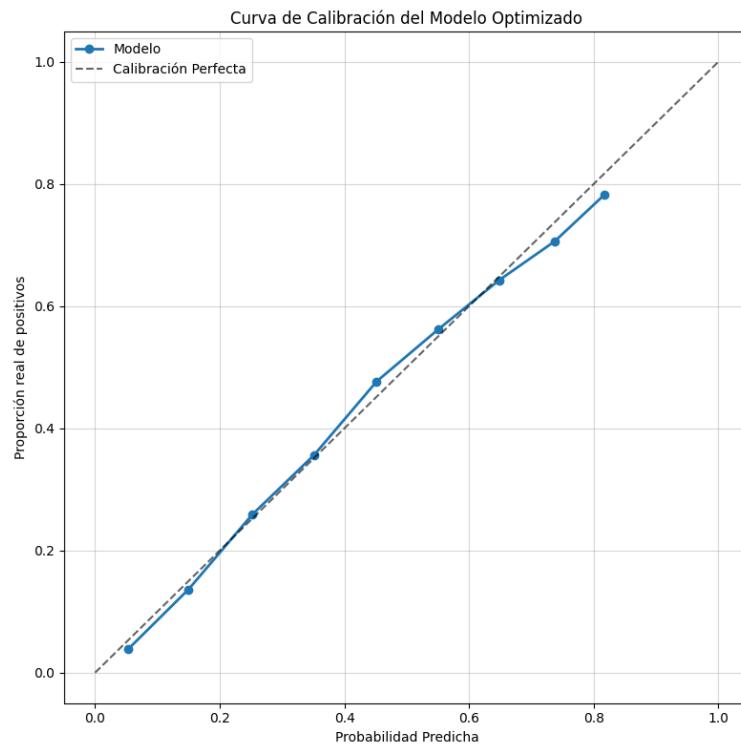


Figura 10: Curva de calibración

5. Resultados para la carrera de física

Con el objetivo de que este proyecto trascienda el ámbito individual y genere un impacto académico, se elaboró el presente reporte técnico con una descripción detallada del desarrollo metodológico, el análisis comparativo entre los distintos modelos implementados y una reflexión crítica sobre las limitaciones encontradas y posibles líneas de trabajo futuro. Asimismo, se buscó que las conclusiones obtenidas sean de utilidad para proyectos con flujos de trabajo similares o con objetivos compartidos dentro del análisis de datos en física de altas energías. El código desarrollado junto con los conjuntos de datos, las transformaciones aplicadas y las visualizaciones de desempeño, se encuentran disponibles en un repositorio de acceso abierto:

https://github.com/marianophys/ServicioSocial_FCiencias_Fisica

El proyecto aporta una herramienta reproducible que puede ser empleada en asignaturas como Temas Selectos de Física Computacional, Física Computacional o áreas afines al análisis de datos del CERN, promoviendo la integración de técnicas modernas de Inteligencia Artificial en el estudio de datos reales de física de partículas. De esta manera, se fortalece la formación en análisis cuantitativo y modelado computacional dentro del contexto de la física experimental contemporánea.

6. Discusión y Conclusiones

La aplicación de los filtros físicos tuvo un impacto determinante en la definición del problema. Aunque inicialmente se consideraban cuatro procesos (Higgs, WW, $t\bar{t}$ y $Z \rightarrow \tau\tau$), tras la selección los eventos de $t\bar{t}$ y $Z \rightarrow \tau\tau$ se volvieron marginales. Esto confirma que los cortes aplicados están bien diseñados y permiten focalizar el análisis en el fondo irreducible dominante (WW), simplificando el problema a la distinción entre señal de Higgs y dibosones.

El histograma de la masa invariante del sistema dileptónico m_{LL} (figura 5) muestra que esta variable posee poder discriminante, aunque con un solapamiento considerable entre señal y fondo. Esto refleja una limitación física intrínseca: las diferencias cinemáticas entre ambos procesos son sutiles y no permiten una separación clara mediante una sola variable.

Come se mención, flujo experimental del proyecto se estructuró en tres etapas de entrenamiento de modelos: (i) entrenamiento de modelos base sin procesamiento adicional, (ii) entrenamiento con una selección de características y (iii) optimización bayesiana del modelo LightGBM. Sin embargo, al comparar cuantitativamente las métricas de evaluación obtenidas en cada fase, se observa que las mejoras alcanzadas son marginales.

De hecho aunque variables como la masa invariante dileptónica muestran poder discriminante, la separabilidad entre clases es intrínsecamente limitada por la similitud ci-

nemática entre señal y fondo. En consecuencia, todos los modelos evaluados ,independientemente de su complejidad, convergen hacia un desempeño similar.

En el caso de los modelos base (Tabla 1), todas las arquitecturas, aun perteneciendo a familias conceptualmente distintas (modelos lineales, ensambles tipo bagging y boosting, y métodos basados en márgenes), presentan desempeños muy similares, con valores de AUC en el rango [0.732, 0.747] y variaciones en accuracy del orden de milésimas. Esta tendencia se mantiene tras la selección de características (Tabla 3), donde las diferencias siguen siendo estadísticamente pequeñas y dentro de las desviaciones estándar reportadas.

La optimización bayesiana aplicada a LightGBM (Tabla 4) produce una mejora ligera en AUC-ROC, alcanzando 0.749 ± 0.005 , pero el incremento absoluto respecto al mejor modelo base es mínimo. Considerando el costo computacional asociado a la búsqueda de hiperparámetros, la ganancia obtenida puede considerarse baja en términos prácticos.

Este comportamiento sugiere que el límite de desempeño no está determinado principalmente por la capacidad expresiva del modelo, sino por la información intrínseca contenida en las variables disponibles. La relativa homogeneidad en el rendimiento entre modelos lineales y no lineales indica que la separabilidad de los datos está condicionada por la estructura física del problema y por las variables reconstruidas, más que por la complejidad algorítmica empleada. No se observan indicios claros de sobreajuste, lo cual refuerza la hipótesis de que el desempeño alcanzado representa un techo práctico bajo el conjunto actual de características.

Las métricas del modelo optimizado confirman una capacidad discriminante moderada: la curva ROC (AUC = 0.749) y la Precision–Recall (AP = 0.645) muestran un desempeño consistente pero no sobresaliente. La matriz de confusión indica que se identifica correctamente el 64 % de la señal y el 71 % del fondo, evidenciando un solapamiento significativo entre clases. La distribución del score refuerza esta observación, mientras que la buena calibración del modelo indica que las probabilidades predichas son confiables y coherentes con las frecuencias reales.

En conjunto, los resultados muestran un modelo sólido, estable y bien calibrado, cuyo desempeño está principalmente limitado por la estructura física del problema y la información contenida en las variables reconstruidas. El proyecto demuestra que la integración rigurosa de fundamentos físicos con herramientas estadísticas modernas es esencial; sin embargo, cualquier mejora significativa futura dependerá más de nuevas variables o representaciones físicas más informativas que de ajustes adicionales del modelo.

Apéndices

A. Diccionario de datos

Como se mencionó previamente, los datos empleados en este proyecto provienen de los Datos Abiertos del CERN, disponibles en [2]. Para el análisis se seleccionó como señal el proceso ilustrado en la Figura 3, correspondiente al decaimiento del bosón de Higgs:

$$ggF \rightarrow H \rightarrow WW^* \rightarrow \ell\nu\ell\nu,$$

el cual constituye la señal que deseamos identificar. Además, se consideraron tres procesos que actúan como fondo:

Dibosón:

$$R \rightarrow W^+W^- \rightarrow \ell\nu\ell\nu,$$

$Z \rightarrow \tau\tau$:

$$Z \rightarrow \tau\tau \rightarrow \ell\nu\ell\nu,$$

$t\bar{t}$:

$$R \rightarrow t\bar{t} \rightarrow W^+W^- b\bar{b} \rightarrow \ell\nu\ell\nu b\bar{b},$$

donde R representa una resonancia producida en la colisión protón–protón.

El objetivo es entrenar un modelo de clasificación capaz de distinguir correctamente los eventos asociados al bosón de Higgs de aquellos generados por estos procesos de fondo.

Descripción de las Variables

Los archivos .root contienen 85 variables en total, de las cuales se seleccionó un subconjunto necesario para aplicar filtros de calidad y para construir el conjunto de entrenamiento. Las definiciones provienen del diccionario de datos oficial de ATLAS Open Data [2] y del documento técnico correspondiente [3]. A continuación se describen las variables seleccionadas, clasificadas según su función en el análisis.

Variables utilizadas para filtrado de eventos

Estas variables permiten aplicar cortes básicos para asegurar que el evento cumple los requisitos mínimos de reconstrucción y calidad establecidos en el análisis del canal $H \rightarrow WW^*$. Es decir, la selección de estas variables se lleva a cabo para asegurar que se cumplen los requerimientos establecidos en el capítulo 3.5 de [3] sirven para asegurar la calidad mínima necesaria de los datos y que estos son candidatos a ser la señal buscada. Aunque no todas las variables poseen un significado físico directo, resultan útiles para caracterizar adecuadamente los eventos y complementar la selección final.

- **trigE, trigM** (booleanas): indican si el evento activó un disparo basado en electrones (`trigE`) o en muones (`trigM`). Son esenciales para garantizar la eficiencia de selección del canal dileptónico.
- **lep_n** (entero): número total de leptones reconstruidos en el evento.
- **jet_n** (entero): número total de jets reconstruidos en el evento.
- **lep_isTightID** (booleana): indica si un leptón cumple con los criterios estrictos de identificación ("Tight ID").
- **lep_type** (entero): tipo de leptón (11 = electrón, 13 = muón).
- **lep_charge** (entero): carga eléctrica del leptón (+1 o -1).
- **jet_MV2c10** (float): valor del discriminante del algoritmo de *b-tagging* MV2c10, que estima la probabilidad de que el jet provenga de un quark *b*. Útil para descartar eventos con múltiples jets *b*, característicos del fondo $t\bar{t}$.

Variables utilizadas para el entrenamiento del modelo

Estas variables contienen información cinemática y de identificación que caracteriza el estado físico del evento y permiten distinguir la señal $ggF \rightarrow H \rightarrow WW^*$ de los procesos de fondo.

■ Variables de leptones

- **lep_pt** (float): momento transversal p_T de cada leptón (en MeV).
- **lep_eta** (float): pseudorapidez del leptón.
- **lep_phi** (float): ángulo acimutal del leptón (en radianes).
- **lep_ptcone30** (float): aislamiento basado en la suma de p_T alrededor del leptón dentro de un cono $\Delta R = 0.3$.
- **lep_etcone20** (float): aislamiento basado en energía depositada en calorímetros dentro de un cono $\Delta R = 0.2$.
- **lep_E** (float): energía total reconstruida del leptón.

■ Variables de jets

- **jet_pt** (float): momento transversal de cada jet.
- **jet_eta** (float): pseudorapidez del jet.
- **jet_phi** (float): ángulo acimutal del jet.
- **jet_E** (float): energía total reconstruida del jet.

■ Variables de energía faltante (MET)

- **met_et** (float): energía transversal faltante (MET), asociada a partículas no detectadas como neutrinos.

- **met_phi** (float): dirección acimutal de la MET.

Estas variables permiten reconstruir la cinemática del evento y proporcionan la información necesaria para que el modelo aprenda a diferenciar la señal del proceso $ggF \rightarrow H \rightarrow WW^*$ de los principales fondos, como WW , $Z \rightarrow \tau\tau$ y $t\bar{t}$.

Referencias

- [1] Experimento atlas. https://es.wikipedia.org/wiki/Experimento_ATLAS. Consultado el 2025-11-29.
- [2] ATLAS Collaboration. The 13 tev atlas open data set for education: details. https://opendata.atlas.cern/docs/data/for_education/13TeV_details, 2020. Consultado el 2025-11-30.
- [3] ATLAS Collaboration. Atlas open data: Review of the 13 tev atlas open data release. <https://cds.cern.ch/record/2707171/files/ANA-OTRC-2019-01-PUB-updated.pdf>, 2020. Consultado el 2025-11-29.
- [4] ATLAS Collaboration. The standard model and beyond — atlas open data documentation. https://opendata.atlas.cern/docs/documentation/introduction/SM_and_beyond, 2025. Consultado el 2025-11-29.
- [5] ATLAS Collaboration. Atlas open data: 13 tev proton–proton collision data for education. https://opendata.atlas.cern/docs/data/for_education/13TeV_details, 2026. Sitio consultado: 2026-01-10.
- [6] ATLAS Collaboration. Atlas open data: Hybrid environment setup for 13 tev data analysis. <https://opendata.atlas.cern/docs/13TeV25Doc/enviroments/hybrid>, 2026. Sitio consultado: 2026-01-10.
- [7] Docker Inc. What is a container? <https://www.docker.com/resources/what-container/>, 2026. Sitio consultado: 2026-01-15.
- [8] LHC Higgs Cross Section Working Group, D. de Florian, C. Grojean, F. Maltoni, C. Mariotti, A. Nikitenko, M. Pieri, P. Savard, M. Schumacher, and R. Tanaka. Handbook of lhc higgs cross sections: 4. deciphering the nature of the higgs sector. *CERN Yellow Reports: Monographs*, 2/2017, 2017. CERN-2017-002-M.
- [9] Optuna Contributors. Optuna: A hyperparameter optimization framework. <https://optuna.readthedocs.io/en/stable/index.html>, 2026. Framework para la optimización automática de hiperparámetros. Sitio consultado: 2026-01-15.
- [10] ROOT Team. Root: An object-oriented data analysis framework. <https://root.cern/about/>, 2026. Desarrollado por CERN. Sitio consultado: 2026-01-15.
- [11] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.