

# Detección del Bosón de Higgs

Martín Alejandro Castro Álvarez

Universidad Internacional de Valencia

Calle Pintor Sorolla, 21 46002, Valencia (España)

martinecastro.10.5@gmail.com

Fuente: <https://github.com/MartinCastroAlvarez/higgs-boson-machine-learning>

**Resumen.** El bosón de Higgs es la partícula responsable de darle masa a la materia. Sin embargo, detectar su presencia es un desafío debido a su rareza y rápida desintegración. Este informe propone una estrategia de Deep Learning para poder detectar su presencia en tiempo real gracias a las señales generadas en el LHC del CERN. Un proceso de limpieza de datos, detección de anomalías, reducción de la dimensionalidad, normalización, y visualización, ha revelado una relación multidimensional compleja entre los datos de colisión de partículas y la presencia de esta partícula exótica. Como resultado, se ha implementado una red neuronal secuencial de varias capas por su habilidad para aprender patrones no lineales. El modelo resultante posee una certeza de aproximadamente 77%, y al ajustar el umbral se puede obtener una precisión de aproximadamente un 97%. Los resultados no sólo nos llevan a un mejor entendimiento de las leyes fundamentales de la naturaleza, sino que también prometen llevarnos a nuevos descubrimientos y avances tecnológicos.

**Keywords:** Higgs Boson, Particle Physics, Machine Learning, Large Hadron Collider, CERN, Neural Networks, Deep Learning, Data Preprocessing, Anomaly Detection, Dimensionality Reduction, Data Augmentation, Exploratory Data Analysis, Robust Estimators, Visualization, Multi-Layer Perceptron, Binary Classification, Dropout Layer, Overfitting, Precision, Recall, Accuracy, Cross-Validation, Xavier Initialization, Adam Optimizer, Early Stopping.

## 1. Introducción

### 1.1. Problema

Nosotros y todo lo que nos rodea estamos formados por partículas. Sin embargo, estas partículas no siempre tuvieron masa, y se movían a la velocidad de la luz, por lo que, bajo esas condiciones, el universo como lo conocemos hoy en día no sería posible. Entender el bosón de Higgs, mensajero del campo de Higgs, es el secreto para entender cómo las partículas eventualmente ganaron masa, y permitieron que el universo alcance el estado actual. [1].

### 1.2. Viabilidad

Machine Learning (ML) puede contribuir significativamente al estudio del bosón de Higgs debido a su gran capacidad para manejar, analizar, e interpretar la gran y compleja cantidad de datos producidos por los

colisionadores de partículas. Por lo tanto, pueden facilitar la detección de señales y el estudio de las propiedades del bosón de Higgs, y de otras partículas.

### **1.3. Importancia**

En primer lugar, el volumen de datos generados en los experimentos de física de partículas, como los llevados a cabo en CERN [1], son enormes y complejos. ML puede contribuir en el análisis de datos de manera más eficiente que los métodos tradicionales. En segundo lugar, la rareza y rápida desintegración del bosón de Higgs lo convierten en un candidato perfecto para la capacidad de detección de anomalías que las técnicas de ML modernas nos otorgan.

### **1.4. Impacto**

Al aplicar ML al estudio del bosón de Higgs, podemos lograr nuevos descubrimientos sobre la materia, la energía, y las condiciones del universo en sus orígenes. Además, hay muchísimos avances tecnológicos que se pueden derivar de esta investigación, de la misma manera que ha sucedido con otros descubrimientos en el campo de la física de partículas en medicina, navegación aeroespacial, etc [1].

## **2. Datos**

### **2.1. Fuente**

El dataset, descargado de UC Irvine Machine LEarning Repository [2], y referenciado por Kaggle [3], presenta el problema de clasificar 11 millones de eventos de colisiones de partículas en 2 categorías: aquellos en los que se ha producido un Bosón de Higgs, y aquellos en los que no. Cada evento ha sido obtenido de una colisión de partículas simulada según lo observado por el Large Hadron Collider (LHC) en el CERN. Cada evento está representado por 28 atributos que describen la trayectoria y la desintegración de partículas resultantes. Estos atributos presentan los valores cinemáticos de baja energía medidos en los detectores de partículas, y propiedades de alta energía obtenidas matemáticamente por físicos expertos.

### **2.2. Limpieza de Datos**

De acuerdo a las prácticas comunes de Machine Learning [4], durante el proceso de limpieza de datos, se ha verificado que los atributos son de tipo numérico. Por lo tanto, no es necesario realizar conversión de tipo de datos. Luego, aquellos eventos que se encontraban duplicados han sido eliminados, del mismo modo que aquellos eventos en los cuales faltaba algún dato. Finalmente, los datos han sido normalizados (Ver Fig. 1), para asegurar que cada atributo contribuya de forma uniforme a la detección del bosón de Higgs.

### **2.3. Detección de Anomalías**

Las anomalías, definidas como aquellas que se encuentran más allá de 3 desviaciones estándar de la media (Ver Fig. 2), podrían resultar en que una cantidad significativa de datos sean eliminados. Por lo tanto, se ha optado por definir una anomalía como aquella que se encuentra más allá del percentil 99. De esta manera,

se garantiza que sólo las medicinas más extremas, causadas por error de medición o por ruido aleatorio, sean eliminadas del proceso de entrenamiento.

## 2.4. Reducción de Dimensionalidad

La técnica para reducir la dimensión del dataset consiste en presumir que todos los atributos contribuyen a la detección del bosón de Higgs, pero descartar aquellos que, según el valor R-squared [5], tienen menor poder explicativo de la varianza de la señal que indican su presencia. De este modo, sólo la mejor combinación de atributos tiene que ser utilizada en la fase de entrenamiento (Ver Fig. 3), y se evita el problema de la explosión combinatoria [6].

## 2.5. Data Augmentation

Se ha descartado la necesidad de utilizar un modelo generativo ya que este dataset cuenta con una distribución balanceada entre los casos positivos y negativos, como se demuestra en la fase de exploración de datos. Además, se ha llevado a cabo un test de hipótesis estadístico para demostrar que el hecho de contar con 5.8 millones de casos positivos y 4.8 millones de casos negativos no permiten descartar la idea de que los datos se encuentran distribuidos de forma aleatoria y balanceada; y el hecho de que la relación no sea 50% se debe simplemente al proceso de muestreo, con un 10% de margen de error.

# 3. Exploración

## 3.1. Intuición de Campo

El diagrama de Feynman (Ver Fig. 4) describe el proceso mediante el cual dos gluones se fusionan formando dos bosones de Higgs pesados y eléctricamente neutros, que eventualmente se desintegran en un bosón W y quarks [7]. Desafortunadamente, puede ocurrir un proceso similar en el que las mismas partículas chocan y los resultados luego de la colisión son los mismos, pero sin que se haya generado un bosón de Higgs en el medio.

Las columnas en el dataset corresponden con los siguientes fenómenos físicos, derivados de experimentos reales en el LHC:

1. '**signal**' que es 1 en caso de que la colisión haya dado lugar a un bosón de Higgs, ó 0 en caso contrario.
2. '**lepton pT**' mide el momento de los leptones (como los electrones y los muones).
3. '**lepton eta**' y '**lepton phi**' son los ángulos que describen la dirección, relativa al rayo, en la que los leptones se están moviendo.
4. '**missing energy magnitude**' es la cantidad de energía perdida luego de la colisión, debido a, de acuerdo a la ley de conservación de la energía, partículas que no han sido detectadas.
5. '**missing energy phi**' indica la dirección en la que la energía perdida se puede haber escapado.
6. '**jet 1 pt**', '**jet 2 pt**', '**jet 3 pt**', y '**jet 4 pt**' indica el momento de los jets (haces de partículas que vienen de los quarks ó gluones).
7. '**jet 1 eta**', '**jet 1 phi**', '**jet 2 eta**', '**jet 2 phi**', '**jet 3 eta**', '**jet 3 phi**', '**jet 4 eta**', y '**jet 4 phi**' son los ángulos que describen la dirección en la que se detectaron los jets.

8. '**jet 1 b-tag**', '**jet 2 b-tag**', '**jet 3 b-tag**', '**jet 4 b-tag**' indican si alguno de los jets detectados correspondan a una partícula b-quark pesada.
9. '**m\_jj**', '**m\_jjj**', '**m\_lv**', '**m\_jlv**', '**m\_bb**', '**m\_wbb**', '**m\_wwbb**' son valores de alta energía derivados por físicos, y que pueden ayudar en la detección del bosón de Higgs.

### 3.2 Resumen Estadístico

Datos estadísticos han sido extraídos del dataset, priorizando estimadores robustos [8] que son menos sensibles a anomalías. Por cada atributo, los siguientes estadísticos robustos de ubicación fueron calculados (Ver Fig. 5): Mínimo, máximo, deciles, mediana, trimmed mean, y winsorized mean. Además, los siguientes estadísticos robustos de dispersión (Ver Fig. 6): varianza, desvío estándar, skewness, kurtosis, rango intercuartil.

### 3.3. Visualización

Se han generado, por un lado, histogramas apilados por cada atributo, agrupados por aquellos valores en los que hay bosón de Higgs y aquellos en los que no (Ver Fig. 7). Dado que el atributo a predecir es de tipo booleano, los histogramas son una herramienta más poderosa que los scatter plots y heatmaps en este caso. Como resultado, se puede apreciar que los datos se encuentran lo suficientemente distribuidos simétricamente, por lo que el dataset es apropiado para la etapa de entrenamiento.

Dado que algunos histogramas se encuentran levemente alineados a la izquierda (tal y como indican las métricas de skewness y kurtosis), los atributos fueron centrados (Ver Fig. 8). Lamentablemente, el dataset centrado fué luego descartado durante el entrenamiento debido a sus pobres resultados. La métrica accuracy se mantuvo por debajo del 50% hasta que esta decisión fué tomada.

Finalmente, se han generado gráficos de líneas indicando la acumulación de datos por decil (Ver Fig. 9). Estos gráficos confirman que los datos se encuentran bien distribuidos y no hace falta más limpieza de datos.

Cabe destacar que tanto los histogramas como el gráfico de distribución de deciles han contribuido en la limpieza de datos, guiando, específicamente, el proceso de normalización.

### 3.4. Hypothesis

De acuerdo a las observaciones de las secciones 3.1, 3.2, and 3.3, los datos sugieren que:

1. El atributo '**lepton pt**' se encuentran sesgados hacia la izquierda, indicando que el momento de los leptones tiende a ser bajo. De acuerdo a 3.1, es el comportamiento esperado, ya que las partículas son más raras entre que más energía poseen.
2. El atributo '**lepton phi**' se encuentra uniformemente distribuido, tal y como indica la simetría de su histograma. Esto significa que la dirección del leptón es completamente aleatoria.
3. Todos los atributos '**jet b-tag**' presentan un comportamiento casi booleano, bastante alineado a las conclusiones de 3.1.

4. El atributo '**missing energy magnitude**' y, en general, todos los atributos '**phi**' parecen no influir de ninguna manera en la señal del bosón de Higgs.
5. Todos los atributos '**pt**' muestran una distribución de cola larga, indicando que las partículas luego de la colisión también tienen mayor probabilidad de tener un momento de baja energía. Esta información podría contener el secreto para detectar el bosón de Higgs.
6. Los atributos '**eta**' se encuentran centrados. Esto indica que los ángulos resultantes tienen una distribución normal.
7. Los atributos '**phi**' se encuentran uniformemente distribuidos, indicando que los ángulos resultantes son completamente aleatorios.
8. Los atributos '**m**' también poseen una distribución de cola larga. Esto indica que los atributos de mayor energía tienen baja probabilidad de ocurrir.
9. En los atributos '**jet b-tag**' hay consistentemente mayor cantidad de instancia en las que el bosón de Higgs está presente.
10. Crear gráficos en 3 dimensiones (o incluso en más dimensiones) para realizar un análisis multidimensional es impráctico debido a la cantidad exponencial de combinaciones.

## 4. Optimización

### 4.1. Selección del Modelo

El análisis previo ha revelado que las relaciones multidimensionales del dataset limitan la efectividad de los modelos lineales, tales como la regresión lineal. Además, mientras que los árboles de decisión y los random forests ofrecen mayor adaptación a las relaciones no lineales, requieren un método sesgado para separar los datos en ramas, tales como la entropía de Shannon. Como resultado, un método más flexible es necesario. Las redes neuronales, también conocidas como "aproximadores universales" [9] son preferidas.

### 4.2. Aprendizaje Supervisado

El dataset ya ha sido etiquetado. Es decir, aquellos eventos en los que hay presencia del bosón de Higgs han sido adecuadamente identificados, y diferenciados de aquellos en los que no. Por lo tanto, un método de aprendizaje no supervisado es preferido.

### 4.3. Arquitectura de la Red Neuronal

Se ha utilizado un perceptrón de muchas capas (MLP) (Ver Fig. 10), debido a su capacidad para capturar patrones complejos [10]. Cada capa corresponde a un conjunto de perceptrones descritos por (1), que posee una entrada  $x$ , la matriz de pesos  $w$ , y el sesgo  $b$ .

Una MLP con muchas capas ocultas y una función de activación ReLU (2) ofrecen la capacidad de lidiar con semejante complejidad, al mismo tiempo que evitan el sobreajuste. Sin embargo, la función de activación de la última capa es la función sigmoid (3), de modo tal que los resultados quedan comprendidos entre 0 y 1.

$$z = wx + b \quad (1)$$

$$a = \max(0, z) \quad (2)$$

$$a = 1 / [1 + \exp(-z)] \quad (3)$$

Las capas ocultas de la red neuronal permiten que el modelo aprenda patrones que, de otro modo, permanecen ocultos en los atributos de entrada, logrando así generalizar.

Además, las capas de dropout evitan que el modelo le otorgue demasiada importancia a algunos atributos. De este modo, el modelo se ve forzado a descartar aleatoriamente (aunque con muy baja probabilidad) algunas entradas, para que el modelo se enfoque en aprender patrones.

La función binary cross entropy (4) ha sido utilizada como función de costo debido a que es recomendada [10] para los clasificadores binarios. Esta función recompensa verdaderos positivos ( $y = 1, a = 1$ ), y verdaderos negativos ( $1 - y = 1$  y  $1 - a = 1$ ), al mismo tiempo que castiga falsos positivos ( $y = 0$  y  $a = 1$ ) y falsos negativos ( $y = 1$  y  $a = 0$ ). Sin embargo, dado que esta función se calcula por evento, es necesario obtener un promedio según (5) para tener un mejor estimador del error de clasificación.

$$L(y, a) = y * \log(a) + (1 - y) * \log(1 - a) \quad (4)$$

$$J = -1/N * \sum L(y, a) \quad (5)$$

Las capas convolucionales y los filtros no son requeridos, al tratarse de tan sólo 25 atributos [10]. Dichas capas son recomendadas para casos de alta dimensionalidad, como en el caso de las imágenes.

El optimizador de Adam ha sido elegido debido a su computación automática de learning rates, evitando así que sea necesario ajustar otro hiper-parámetro. Además, otros optimizadores tales como RMSProp y Adagrad han sido utilizados, pero descartados debido a su pobre contribución al accuracy.

The Adam optimizer was chosen due to its automatic computation of adaptive learning rates, to avoid having to adjust the learning rate, although RMSProp and Adagrad have also been tested.

Finalmente, el método de Xavier (Glorot) ha sido utilizado para inicializar los parámetros de las matrices de pesos aleatoriamente. De este modo, se evita el problema de la convergencia lenta.

#### 4.4. Data Splitting

El dataset ha sido dividido en el dataset de entrenamiento, validación, y prueba, de acuerdo a [9]. El dataset de entrenamiento ha sido utilizado para entrenar el modelo, mientras que el dataset de validación ha sido utilizado en la fase de hyper-parameter tuning. Por otro lado, el dataset de prueba ha sido utilizado para obtener una evaluación no sesgada de las métricas del modelo, que indiquen si ha habido overfitting. Además, se ha utilizado una estrategia de muestreo estratificado para mantener una distribución uniforme

de señales positivas y negativas de la presencia del bosón de Higgs en los 3 datasets. De este modo, se evita entrenar o evaluar al modelo en una única categoría.

## 5. Resultados

### 5.1. Hyper-Parameter Tuning

La arquitectura del modelo ha mutado muchas veces, en base a los resultados de las métricas obtenidas durante las pruebas (Ver Fig. 11). Inicialmente, el modelo había sido entrenado con menos de 100 neuronas por cada capa. Sin embargo, la baja exactitud del modelo ha demostrado que aquello había sido una estrategia muy pobre para detectar los patrones complejos de los datos. Como resultado, el número de neuronas ha sido aumentado por encima de 300, logrando instantáneamente mejorar la exactitud del model desde 40% hasta aproximadamente 60%.

Además, el modelo había sido originalmente diseñado con un total de 3 capas, de acuerdo lo sugerido en el paper [7]. Desafortunadamente, la exactitud del modelo se mantuvo por debajo de 75% hasta que se agregaron más capas.

El sobreajuste ha sido un problema recurrente. Esto podía ser visto por una diferencia notable entre las métricas sobre el dataset de entrenamiento y el dataset de validación. Como resultado, se han introducido capas de Dropout, para que el modelo deba aprender patrones en lugar de memorizar casos particulares.

Por otra parte, como resultado del proceso de prueba y error, se ha encontrado un balance entre el número de capas, y el tamaño de cada capa, logrando que el entrenamiento se reduzca de 15 minutos a 5 por epoch pero sin sacrificar accuracy.

El mismo proceso de prueba y error ha revelado que mientras que el tamaño de cada batch era muy grande (excediendo 5.000) reducía el tiempo de entrenamiento, también podía dañar la exactitud del modelo en hasta un 10%. A pesar de que Grid Search ha sido utilizado, fué descartado por el costo computacional, que no justificaba un beneficio mayor al de prueba y error.

### 5.2. Métricas de Evaluación

Al final de cada epoch, el set de validación ha sido utilizado para calcular 3 métricas: accuracy ó exactitud, precision y recall. En particular, precision y recall han sido utilizados para evaluar que cada epoch mejoraba al menos una de ellas. Sin embargo, en este problema en particular, no hay ninguna diferencia entre los falsos positivos y los falsos negativos, como ocurre, por ejemplo, en medicina, donde los falsos negativos son fatales. En este caso, ambos son igualmente incorrectos.

Por otro lado, la evolución de la exactitud en el tiempo (Ver Fig. 12) indica que durante los primeros epochs, el modelo fallaba en la tarea de generalizar. Sin embargo, más tarde, comenzó a mostrar signos de overfitting, por lo que se ha incorporado un mecanismo de early stopping.

Además, el uso de un test de prueba, que no ha participado de ninguna manera durante el proceso de entrenamiento, ha arrojado una medición imparcial de 77% de exactitud del modelo.

Como ha sido mencionado, precision y recall han sido estimados, pero la misma información puede obtenerse de una forma más amigable a través de la matriz de confusión (Ver Fig. 13): Aproximadamente 9% de falsos positivos y 13% de falsos negativos.

### 5.3. Implicaciones

La validación cruzada ha sido descartada debido a problemas de rendimiento; Los recursos computacionales requeridos para entrenar y evaluar el modelo utilizando diferentes rebanadas del dataset han sido demasiado altos. Por lo tanto, se sugiere realizar la misma prueba, en el futuro, con un sistema más potente, que posiblemente utilice GPU.

Otras mejores posibles podrían ser implementar estructuras más complejas de redes neuronales, como transformers [11]. Además, se podrían captar más señales en el LHC, involucrando a una mayor cantidad de físicos en el proceso.

Este modelo puede ser utilizado para tomar decisiones en tiempo real sobre la presencia del bosón de Higgs, contribuyendo significativamente a la investigación en el campo de física de partículas, permitiendo realizar nuevos descubrimientos, e innovaciones tecnológicas.

Sorprendentemente, cuando el umbral de decisión se reduce de 0.5 a 0.05, la precisión del modelo aumenta hasta el 97% aproximadamente. De esta manera, incluso aún cuando el modelo raramente produzca una respuesta positiva, cuando lo hace, podemos estar casi seguros de que nos encontramos ante la presencia de esta partícula exótica. Por lo tanto, podría utilizarse en los colisionadores de partículas, donde millones de eventos son generados, para desviar el bosón de Higgs mediante un campo magnético, por un camino distinto al del resto de las partículas, donde sus propiedades pueden ser estudiadas apropiadamente.

## 6. References

- [1] CERN. (n.d.). The Higgs Boson. Home.CERN. <https://home.cern/science/physics/higgs-boson>.
- [2] Whiteson,Daniel. (2014). HIGGS. UCI Machine Learning Repository.  
<https://doi.org/10.24432/C5V312>.
- [3] sldsrt. (2020). Higgs Boson Detection. Kaggle. <https://kaggle.com/competitions/higgs-boson-detection>.
- [4] Aggarwal, C. C. (2016). Recommender Systems: The Textbook. Springer.
- [5] Selvin, S. (1995). Practical Biostatistical Methods.
- [6] Sutton, R. S., & Barto, A. G. (Second edition). Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning series).
- [7] Baldi, P., P. Sadowski, and D. Whiteson (2014). Searching for Exotic Particles in High-energy Physics with Deep Learning. Nature Communications 5. <https://www.nature.com/articles/ncomms5308>
- [8] Jurecková, J. (2019). Methodology in Robust and Nonparametric Statistics. CRC Press.
- [9] Géron, A. (2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow. O'Reilly
- [10] Ng, A. Deep Learning Specialization. Coursera. <https://www.coursera.org/specializations/deep-learning>
- [11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin (2017). Attention is all you need.

## 7. Annex

	Mean	Std	Min	Max		Mean	Std	Min	Max
signal	0.543922	0.498067	0.000000	1.000000	signal	5.439223e-01	0.498067	0.000000	1.000000
lepton pT	0.989767	0.559807	0.274697	5.851532	lepton pT	-2.016273e-17	1.000000	-1.277351	8.684719
lepton eta	-0.000035	1.008226	-2.428158	2.428050	lepton eta	9.819514e-18	1.000000	-2.408313	2.408275
lepton phi	-0.000010	1.006376	-1.742508	1.743236	lepton phi	-1.376352e-17	1.000000	-1.731459	1.732201
missing energy magnitude	0.995501	0.593610	0.000237	6.109808	missing energy magnitude	1.347212e-15	1.000000	-1.676629	8.615599
missing energy phi	0.000202	1.006232	-1.743616	1.743257	missing energy phi	9.337632e-18	1.000000	-1.733018	1.732259
jet 1 pt	0.989880	0.469982	0.137502	4.926362	jet 1 pt	-1.761214e-16	1.000000	-1.813639	8.375816
jet 1 eta	0.000014	1.008626	-2.961803	2.961752	jet 1 eta	1.537244e-17	1.000000	-2.936487	2.936408
jet 1 phi	0.000084	1.005877	-1.741237	1.741454	jet 1 phi	-3.459995e-17	1.000000	-1.731147	1.731195
jet 1 b-tag	0.999737	1.027822	0.000000	2.173076	jet 1 b-tag	-1.583309e-17	1.000000	-0.972675	1.141579
jet 2 pt	0.991545	0.495126	0.188981	5.573496	jet 2 pt	-1.539385e-16	1.000000	-1.620928	9.254109
jet 2 eta	-0.000161	1.008419	-2.909204	2.909324	jet 2 eta	-2.415120e-17	1.000000	-2.884757	2.885195
jet 2 phi	-0.000101	1.006131	-1.742372	1.743175	jet 2 phi	-9.988107e-18	1.000000	-1.731653	1.732653
jet 2 b-tag	0.998620	1.049301	0.000000	2.214872	jet 2 b-tag	-5.168752e-17	1.000000	-0.951700	1.159107
jet 3 pt	0.991414	0.483898	0.263608	5.525750	jet 3 pt	-1.620890e-16	1.000000	-1.504049	9.370434
jet 3 eta	-0.000091	1.007976	-2.727842	2.727278	jet 3 eta	-9.767742e-18	1.000000	-2.706168	2.705788
jet 3 phi	-0.000035	1.006320	-1.742069	1.742884	jet 3 phi	2.649025e-17	1.000000	-1.731094	1.731973
jet 3 b-tag	0.999263	1.193519	0.000000	2.548224	jet 3 b-tag	5.512508e-17	1.000000	-0.837240	1.297810
jet 4 pt	0.985864	0.502963	0.365354	5.433091	jet 4 pt	-1.263594e-16	1.000000	-1.233707	8.842052
jet 4 eta	-0.000090	1.006859	-2.496432	2.496343	jet 4 eta	-5.705593e-18	1.000000	-2.479336	2.479426
jet 4 phi	0.000024	1.006416	-1.742691	1.743372	jet 4 phi	6.019547e-18	1.000000	-1.731605	1.732234
jet 4 b-tag	1.000507	1.400379	0.000000	3.101961	jet 4 b-tag	5.177846e-17	1.000000	-0.714454	1.500633
m_ijj	1.032058	0.652120	0.075070	13.583668	m_ijj	-2.167807e-18	1.000000	-1.467501	19.247378
m_ijjj	1.023499	0.370092	0.198676	7.271092	m_ijjj	2.055753e-16	1.000000	-2.228695	16.881175
m_lv	1.050284	0.161760	0.083049	3.431373	m_lv	-1.135371e-16	1.000000	-5.979446	14.719879
m_jlv	1.007871	0.388983	0.132006	5.877596	m_jlv	3.980308e-17	1.000000	-2.251676	12.519111

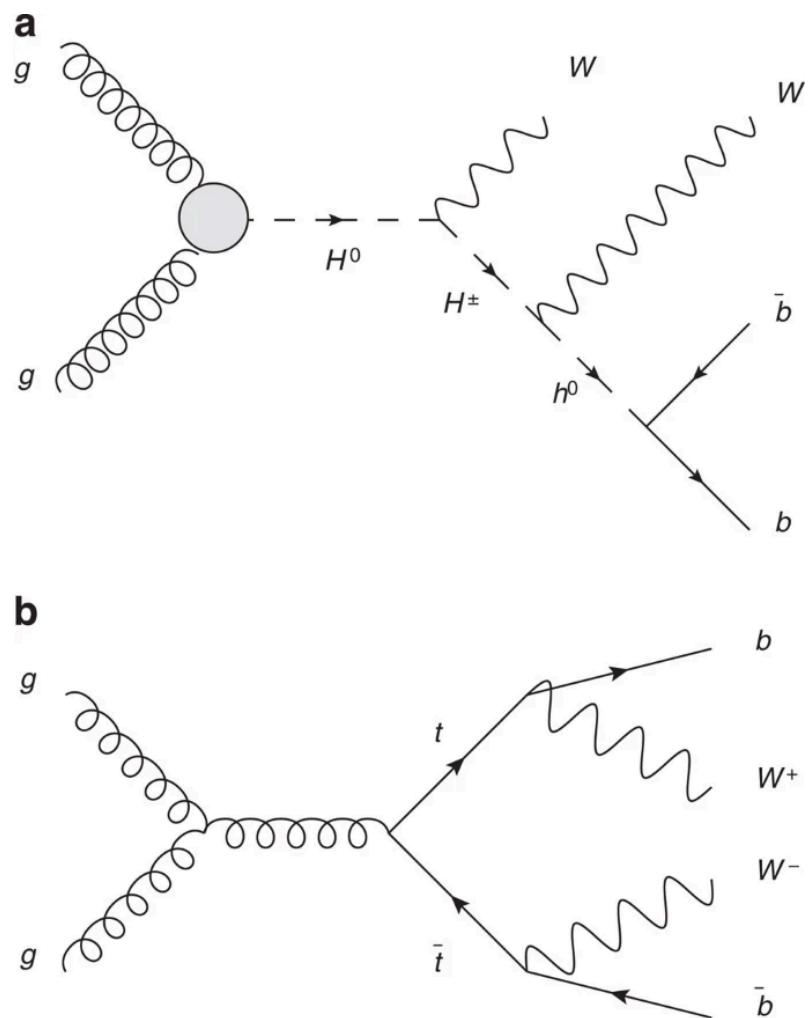
Fig. 1: Resumen estadístico de los datos, antes y después del proceso de normalización.

ignal	0	signal
epton pT	1784	lepton pT
epton eta	0	lepton eta
epton phi	0	lepton phi
issing energy magnitude	1595	missing energy magnitude
issing energy phi	0	missing energy phi
et 1 pt	1333	jet 1 pt
et 1 eta	0	jet 1 eta
et 1 phi	0	jet 1 phi
et 1 b-tag	0	jet 1 b-tag
et 2 pt	2409	jet 2 pt
et 2 eta	0	jet 2 eta
et 2 phi	0	jet 2 phi
et 2 b-tag	0	jet 2 b-tag
et 3 pt	2635	jet 3 pt
et 3 eta	0	jet 3 eta
et 3 phi	0	jet 3 phi
et 3 b-tag	0	jet 3 b-tag
et 4 pt	1915	jet 4 pt
et 4 eta	0	jet 4 eta
et 4 phi	0	jet 4 phi
et 4 b-tag	0	jet 4 b-tag
_jj	23449	m_jj
_jjj	17960	m_jjj
_lv	15054	m_lv
_jlv	8086	m_jlv
_bb	6250	m_bb
_wbb	6800	m_wbb
_wwbb	5589	m_wwbb
		dtype: int64
type: int64		

**Fig. 2:** A la izquierda la cantidad de casos en los que el dato se encuentra a más de 8 desviaciones estándar de la media. A la derecha, los que se encuentran por encima del percentil 99.

signal	0	signal	0
lepton pT	1784	lepton pT	1073
lepton eta	0	lepton eta	1058
lepton phi	0	lepton phi	0
missing energy magnitude	1595	missing energy magnitude	1073
missing energy phi	0	missing energy phi	1063
jet 1 pt	1333	jet 1 pt	1073
jet 1 eta	0	jet 1 eta	1050
jet 1 phi	0	jet 1 phi	0
jet 1 b-tag	0	jet 1 b-tag	0
jet 2 pt	2409	jet 2 pt	1073
jet 2 eta	0	jet 2 eta	1021
jet 2 phi	0	jet 2 phi	0
jet 2 b-tag	0	jet 2 b-tag	0
jet 3 pt	2635	jet 3 pt	1073
jet 3 eta	0	jet 3 eta	991
jet 3 phi	0	jet 3 phi	0
jet 3 b-tag	0	jet 3 b-tag	0
jet 4 pt	1915	jet 4 pt	1073
jet 4 eta	0	jet 4 eta	987
jet 4 phi	0	jet 4 phi	0
jet 4 b-tag	0	jet 4 b-tag	0
m_jj	23449	m_jj	1073
m_jjj	17960	m_jjj	1073
m_lv	15054	m_lv	1073
m_jlv	8086	m_jlv	1073
m_bb	6250	m_bb	1073
m_wbb	6800	m_wbb	1073
m_wwbb	5589	m_wwbb	1073
dtype: int64		dtype: int64	

**Fig. 3:** La imagen a la izquierda muestra el valor R-squared cuando se elimina cada uno de los features. A la derecha, el mismo proceso, pero una vez que se ha eliminado, en el paso anterior, el feature con menor R-squared. A la izquierda se elimina el feature 'm\_bb' removal, mientras que a la derecha se elimina 'm\_wwbb' de acuerdo a [7]



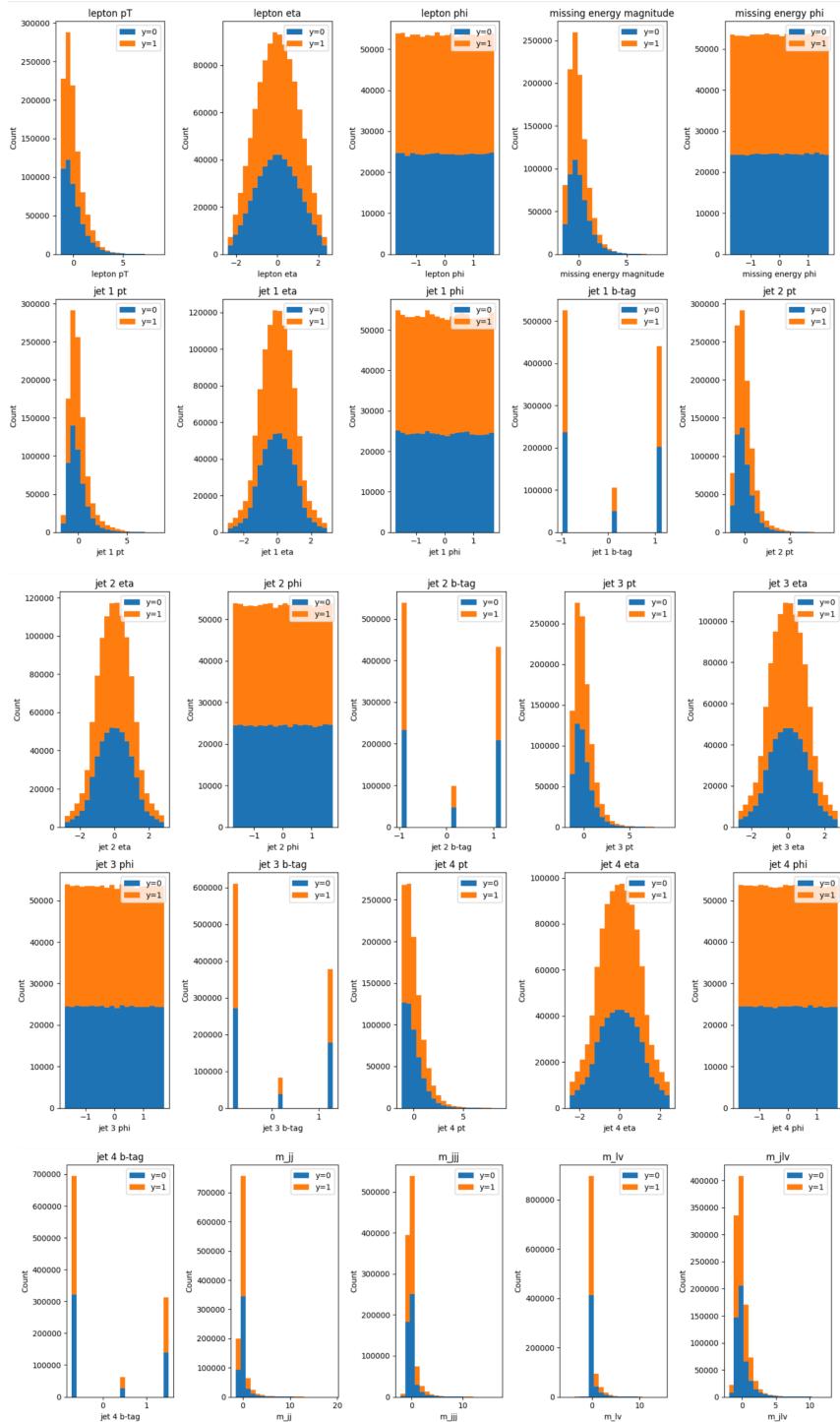
**Fig. 4:** (a) Diagrama de Feynman describiendo la interacción entre partículas en la que participa el bosón de Higgs. (b) Diagrama de Feynman describiendo la interacción que posee las mismas condiciones iniciales y finales pero sin participación del bosón de Higgs. Source: [7]

	max	min	10%	20%	30%	40%	50%	60%	70%	80%	90%
<b>lepton pT</b>	8.684719	-1.277351	-0.997839	-0.802671	-0.622868	-0.440450	-0.244301	-0.019711	0.259147	0.658964	1.333062
<b>lepton eta</b>	2.408275	-2.408313	-1.340859	-0.907115	-0.572871	-0.278234	-0.000019	0.278195	0.571866	0.906110	1.341786
<b>lepton phi</b>	1.732201	-1.731459	-1.385765	-1.039520	-0.692723	-0.346476	-0.000230	0.346117	0.692914	1.039159	1.385956
<b>missing energy magnitude</b>	8.615599	-1.676629	-1.082155	-0.819424	-0.599072	-0.390506	-0.177399	0.055410	0.330454	0.693067	1.281261
<b>missing energy phi</b>	1.732259	-1.733018	-1.385949	-1.039112	-0.692818	-0.346235	0.000152	0.346575	0.692553	1.039054	1.385400
<b>jet 1 pt</b>	8.375816	-1.813639	-0.991481	-0.759336	-0.566368	-0.385486	-0.201485	-0.000915	0.237467	0.563758	1.183592
<b>jet 1 eta</b>	2.936408	-2.936487	-1.246876	-0.841409	-0.537063	-0.257261	-0.000039	0.256200	0.536984	0.841330	1.246797
<b>jet 1 phi</b>	1.731195	-1.731147	-1.388891	-1.037818	-0.691153	-0.349446	0.000077	0.349497	0.691201	1.037866	1.388388
<b>jet 1 b-tag</b>	1.141579	-0.972675	-0.972675	-0.972675	-0.972675	-0.972675	0.084452	1.141579	1.141579	1.141579	1.141579
<b>jet 2 pt</b>	9.254109	-1.620928	-0.998437	-0.773902	-0.583697	-0.398322	-0.204811	0.010569	0.267652	0.612463	1.200626
<b>jet 2 eta</b>	2.885195	-2.884757	-1.254914	-0.845527	-0.539209	-0.261789	0.000219	0.261264	0.538684	0.845965	1.254389
<b>jet 2 phi</b>	1.732653	-1.731653	-1.386446	-1.039033	-0.691621	-0.346962	0.000450	0.346862	0.692069	1.038379	1.386343
<b>jet 2 b-tag</b>	1.159107	-0.951700	-0.951700	-0.951700	-0.951700	-0.951700	-0.951700	1.159107	1.159107	1.159107	1.159107
<b>jet 3 pt</b>	9.370434	-1.504049	-0.1049660	-0.809037	-0.602893	-0.402918	-0.194959	0.035502	0.310240	0.672807	1.265110
<b>jet 3 eta</b>	2.705788	-2.706168	-1.264906	-0.854923	-0.545178	-0.266137	0.000262	0.265757	0.545702	0.855447	1.264526
<b>jet 3 phi</b>	1.731973	-1.731094	-1.385459	-1.039273	-0.692536	-0.347450	-0.000713	0.346678	0.692312	1.039601	1.385787
<b>jet 3 b-tag</b>	1.297810	-0.837240	-0.837240	-0.837240	-0.837240	-0.837240	-0.837240	0.230285	1.297810	1.297810	1.297810
<b>jet 4 pt</b>	8.842052	-1.233707	-1.017867	-0.825257	-0.637486	-0.443423	-0.233391	0.004711	0.292659	0.676428	1.302171
<b>jet 4 eta</b>	2.479426	-2.479336	-1.285759	-0.869703	-0.557867	-0.274155	0.000459	0.274245	0.557958	0.869793	1.285849
<b>jet 4 phi</b>	1.732234	-1.731605	-1.385342	-1.039630	-0.692815	-0.346550	-0.000287	0.346079	0.692893	1.039156	1.385971
<b>jet 4 b-tag</b>	1.500633	-0.714454	-0.714454	-0.714454	-0.714454	-0.714454	-0.714454	-0.714454	0.393089	1.500633	1.500633
<b>m_jj</b>	19.247378	-1.467501	-0.592056	-0.420924	-0.329867	-0.265405	-0.210225	-0.156817	-0.077798	0.095904	0.657581
<b>m_jjj</b>	16.881175	-2.228695	-0.760508	-0.555369	-0.414229	-0.300622	-0.196158	-0.085288	0.063958	0.289803	0.787000
<b>m_lv</b>	14.719879	-5.979446	-0.425828	-0.405540	-0.393544	-0.384971	-0.374045	-0.356749	-0.318345	0.112783	0.994902
<b>m_jlv</b>	12.519111	-2.251676	-0.926343	-0.709802	-0.536446	-0.385585	-0.234802	-0.053912	0.187743	0.540158	1.147213

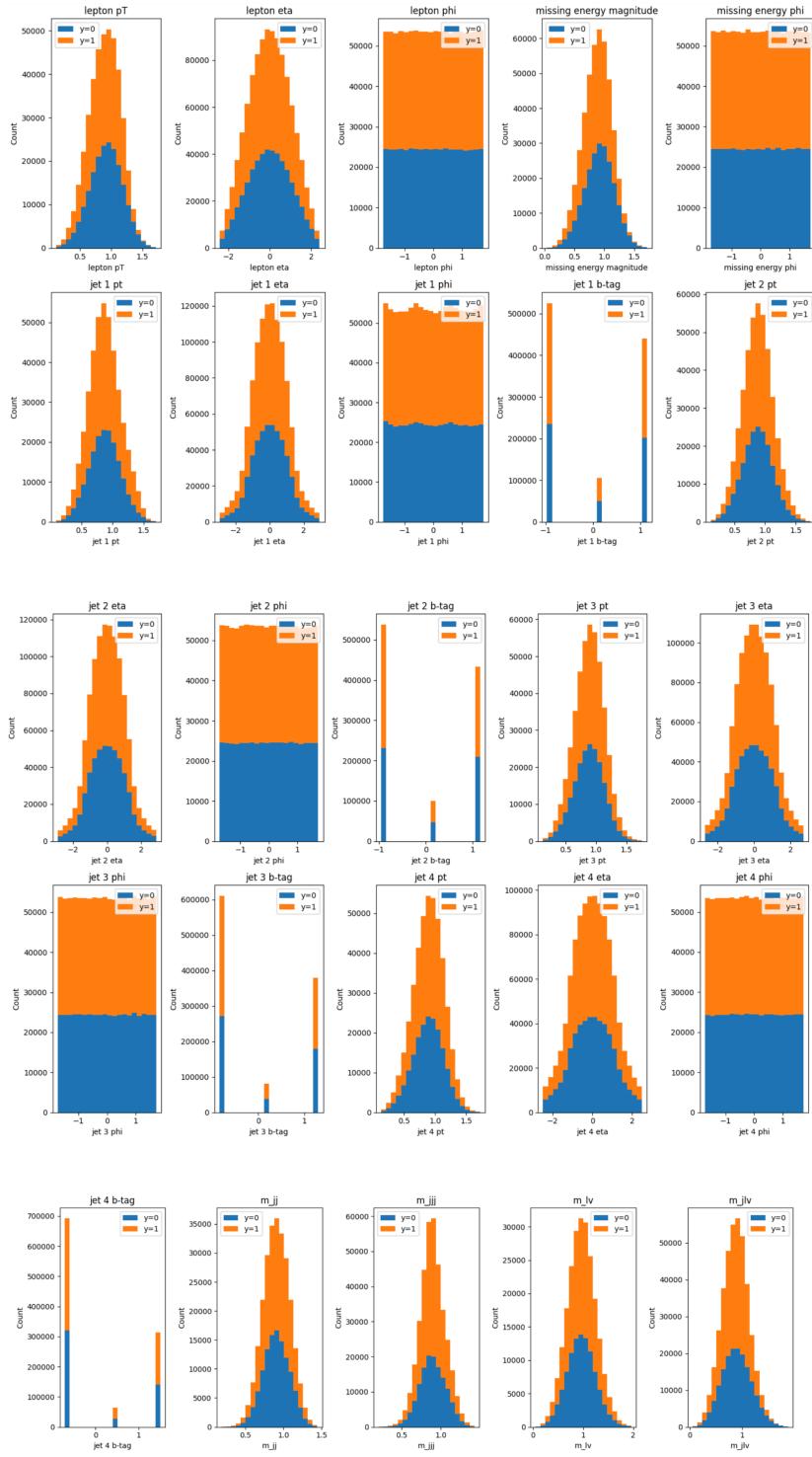
**Fig. 5:** Máximo, mínimo, y deciles de 10% a 90%.

	skew	kurt	median	iqr	trimmed_mean	winsorized_mean
lepton pT	1.662314	4.446112	-0.244301	1.149432	-0.137731	-4.096970e-02
lepton eta	-0.000086	-0.633241	-0.000019	1.464490	-0.000016	1.400163e-05
lepton phi	0.000428	-1.199795	-0.000230	1.731883	-0.000073	4.945425e-06
missing energy magnitude	1.389330	3.695907	-0.177399	1.202787	-0.106226	-3.689057e-02
missing energy phi	-0.000661	-1.199486	0.000152	1.731797	0.000101	6.967534e-06
jet 1 pt	1.803507	5.169810	-0.201485	1.044168	-0.134105	-4.641937e-02
jet 1 eta	0.000200	-0.015604	-0.000039	1.360721	-0.000018	-1.179883e-05
jet 1 phi	-0.000448	-1.200457	0.000077	1.726263	0.000047	1.019412e-05
jet 1 b-tag	0.159581	-1.863448	0.084452	2.114254	-0.021113	-1.583309e-17
jet 2 pt	1.845552	6.134492	-0.204811	1.100803	-0.127195	-4.688541e-02
jet 2 eta	0.000418	-0.053201	0.000219	1.376504	-0.000022	-5.902080e-05
jet 2 phi	0.000114	-1.199231	0.000450	1.729448	-0.000040	-1.997072e-06
jet 2 b-tag	0.196713	-1.857187	-0.951700	2.110807	-0.025926	-5.168752e-17
jet 3 pt	1.586724	4.870944	-0.194959	1.179887	-0.116219	-4.065963e-02
jet 3 eta	-0.000720	-0.196469	0.000262	1.387079	0.000119	3.271479e-05
jet 3 phi	0.000715	-1.200617	-0.000713	1.731585	-0.000082	-1.455435e-05
jet 3 b-tag	0.440632	-1.715070	-0.837240	2.135051	-0.057571	5.512508e-17
jet 4 pt	1.630674	4.418853	-0.233391	1.199702	-0.134758	-4.305734e-02
jet 4 eta	-0.000173	-0.395908	0.000459	1.416907	0.000037	-4.123661e-05
jet 4 phi	0.000250	-1.199747	-0.000287	1.731972	-0.000068	-2.824010e-05
jet 4 b-tag	0.757824	-1.343227	-0.714454	2.215088	-0.098272	5.177846e-17
m_ij	5.552663	49.161982	-0.210225	0.358661	-0.174510	-8.273098e-02
m_jjj	4.279232	31.546421	-0.196158	0.643401	-0.154376	-7.522295e-02
m_lv	4.190406	26.014672	-0.374045	0.215173	-0.242908	-7.286099e-02
m_jlv	2.516012	12.014491	-0.234802	0.964449	-0.141486	-5.560462e-02

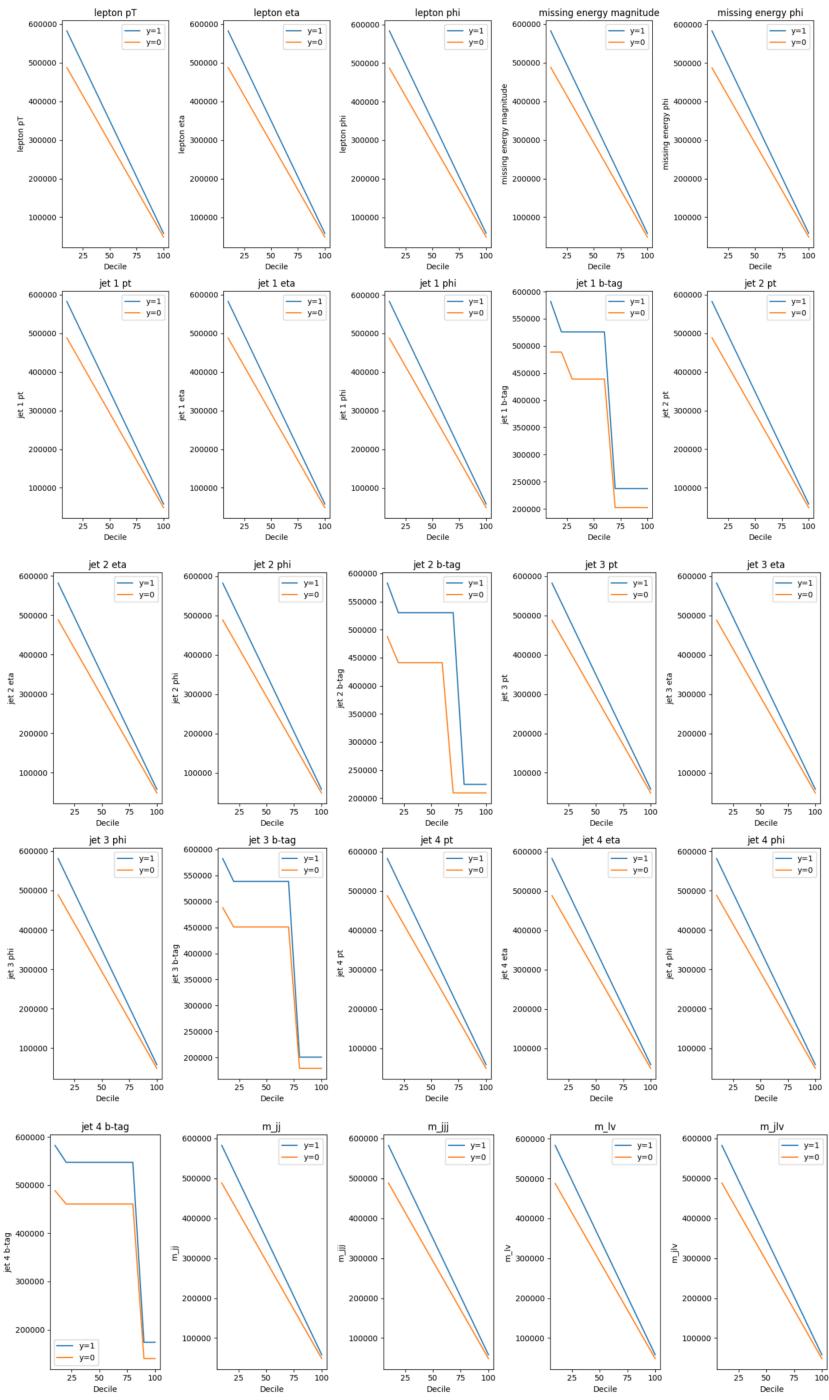
**Fig. 6:** Mediana, rango intercuartil, trimmed mean y winsorized mean.



**Fig. 7:** Histograma de datos, mostrando la concentración de distintos tipos de datos en presencia del bosón de Higgs (naranja) y en su ausencia (azul).



**Fig. 8:** Histograma luego de centrar los datos.



**Fig. 9:** Gráficos de línea indicando la distribución por decil de cada atributo en relación a los casos en los que hay presencia del bosón de Higgs (azul) y los casos en los que no (naranja).

Model: "sequential"

Layer (type)	Output Shape	Param #
<hr/>		
dense (Dense)	(None, 300)	7800
dense_1 (Dense)	(None, 300)	90300
dropout (Dropout)	(None, 300)	0
dense_2 (Dense)	(None, 300)	90300
dropout_1 (Dropout)	(None, 300)	0
dense_3 (Dense)	(None, 300)	90300
dropout_2 (Dropout)	(None, 300)	0
dense_4 (Dense)	(None, 300)	90300
dropout_3 (Dropout)	(None, 300)	0
dense_5 (Dense)	(None, 300)	90300
dropout_4 (Dropout)	(None, 300)	0
dense_6 (Dense)	(None, 300)	90300
dropout_5 (Dropout)	(None, 300)	0
dense_7 (Dense)	(None, 1)	301
<hr/>		
Total params: 549901 (2.10 MB)		
Trainable params: 549901 (2.10 MB)		
Non-trainable params: 0 (0.00 Byte)		

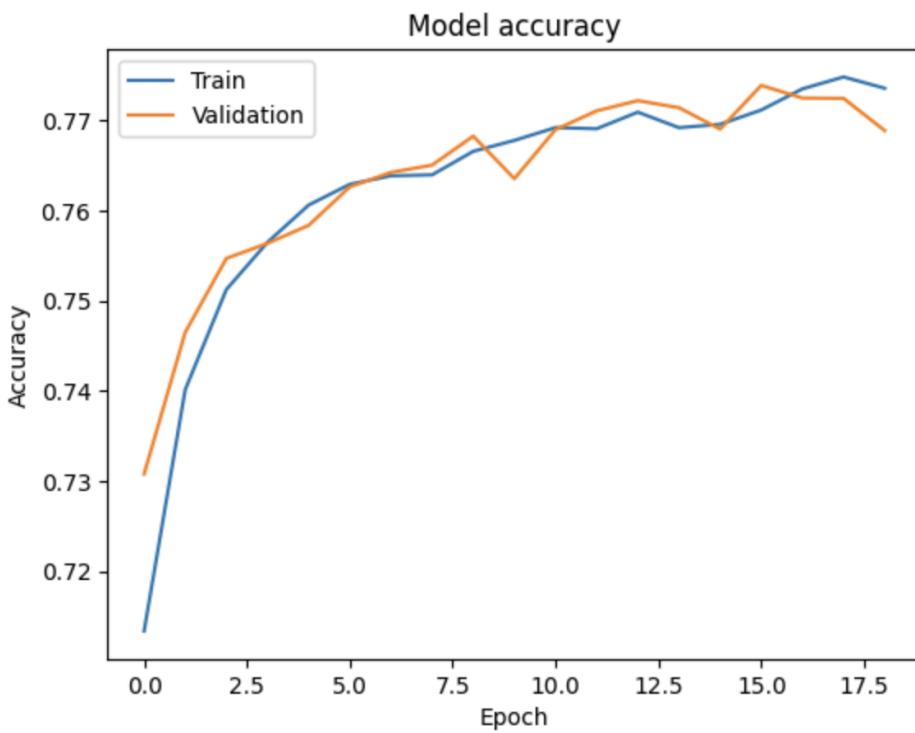
**Fig. 10:** Detalle del modelo secuencial. A la derecha, la cantidad de parámetros a optimizar..

```

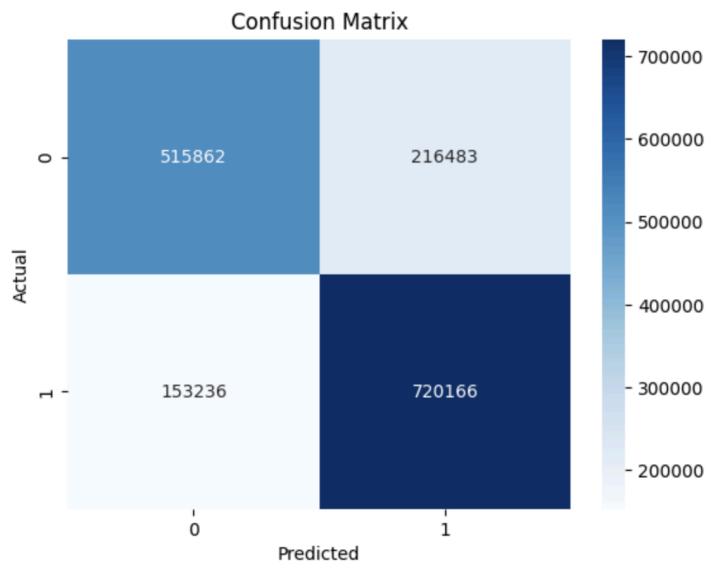
58543/58543 [=====] - 414s 7ms/step - loss: 0.4944 - accuracy: 0.7565 - precision: 0.7604 - recall: 0.8063 - val_loss
s: 0.4937 - val_accuracy: 0.7564 - val_precision: 0.7636 - val_recall: 0.7997
Epoch 5/25
58543/58543 [=====] - 399s 7ms/step - loss: 0.4876 - accuracy: 0.7606 - precision: 0.7639 - recall: 0.8102 - val_loss
s: 0.4879 - val_accuracy: 0.7584 - val_precision: 0.7819 - val_recall: 0.7707
Epoch 6/25
58543/58543 [=====] - 382s 7ms/step - loss: 0.4845 - accuracy: 0.7629 - precision: 0.7667 - recall: 0.8108 - val_loss
s: 0.4835 - val_accuracy: 0.7626 - val_precision: 0.7752 - val_recall: 0.7937
Epoch 7/25
58543/58543 [=====] - 376s 6ms/step - loss: 0.4825 - accuracy: 0.7639 - precision: 0.7659 - recall: 0.8150 - val_loss
s: 0.4821 - val_accuracy: 0.7642 - val_precision: 0.7635 - val_recall: 0.8208
Epoch 8/25
58543/58543 [=====] - 402s 7ms/step - loss: 0.4819 - accuracy: 0.7640 - precision: 0.7675 - recall: 0.8121 - val_loss
s: 0.4809 - val_accuracy: 0.7650 - val_precision: 0.7722 - val_recall: 0.8057
Epoch 9/25
58543/58543 [=====] - 424s 7ms/step - loss: 0.4782 - accuracy: 0.7666 - precision: 0.7696 - recall: 0.8147 - val_loss
s: 0.4752 - val_accuracy: 0.7683 - val_precision: 0.7752 - val_recall: 0.8083
Epoch 10/25
58543/58543 [=====] - 439s 7ms/step - loss: 0.4766 - accuracy: 0.7678 - precision: 0.7703 - recall: 0.8165 - val_loss
s: 0.4821 - val_accuracy: 0.7635 - val_precision: 0.7472 - val_recall: 0.8543
Epoch 11/25
58543/58543 [=====] - 463s 8ms/step - loss: 0.4747 - accuracy: 0.7692 - precision: 0.7715 - recall: 0.8179 - val_loss
s: 0.4747 - val_accuracy: 0.7690 - val_precision: 0.7784 - val_recall: 0.8042
Epoch 12/25
58543/58543 [=====] - 459s 8ms/step - loss: 0.4741 - accuracy: 0.7691 - precision: 0.7724 - recall: 0.8159 - val_loss
s: 0.4712 - val_accuracy: 0.7711 - val_precision: 0.7777 - val_recall: 0.8109
Epoch 13/25
58543/58543 [=====] - 455s 8ms/step - loss: 0.4716 - accuracy: 0.7709 - precision: 0.7736 - recall: 0.8183 - val_loss
s: 0.4692 - val_accuracy: 0.7722 - val_precision: 0.7651 - val_recall: 0.8386
Epoch 14/25
58543/58543 [=====] - 552s 9ms/step - loss: 0.4736 - accuracy: 0.7692 - precision: 0.7721 - recall: 0.8167 - val_loss
s: 0.4710 - val_accuracy: 0.7714 - val_precision: 0.7830 - val_recall: 0.8020
Epoch 15/25
58543/58543 [=====] - 466s 8ms/step - loss: 0.4735 - accuracy: 0.7696 - precision: 0.7717 - recall: 0.8185 - val_loss
s: 0.4741 - val_accuracy: 0.7690 - val_precision: 0.7629 - val_recall: 0.8349
Epoch 16/25
58543/58543 [=====] - 442s 8ms/step - loss: 0.4719 - accuracy: 0.7711 - precision: 0.7735 - recall: 0.8191 - val_loss
s: 0.4677 - val_accuracy: 0.7739 - val_precision: 0.7744 - val_recall: 0.8245
Epoch 17/25
58543/58543 [=====] - 488s 8ms/step - loss: 0.4680 - accuracy: 0.7735 - precision: 0.7757 - recall: 0.8209 - val_loss
s: 0.4713 - val_accuracy: 0.7725 - val_precision: 0.7854 - val_recall: 0.8005
Epoch 18/25
58543/58543 [=====] - 454s 8ms/step - loss: 0.4664 - accuracy: 0.7748 - precision: 0.7766 - recall: 0.8227 - val_loss
s: 0.4698 - val_accuracy: 0.7724 - val_precision: 0.7718 - val_recall: 0.8258
Epoch 19/25
58543/58543 [=====] - 454s 8ms/step - loss: 0.4682 - accuracy: 0.7736 - precision: 0.7759 - recall: 0.8209 - val_loss
s: 0.4761 - val_accuracy: 0.7689 - val_precision: 0.7675 - val_recall: 0.8250

```

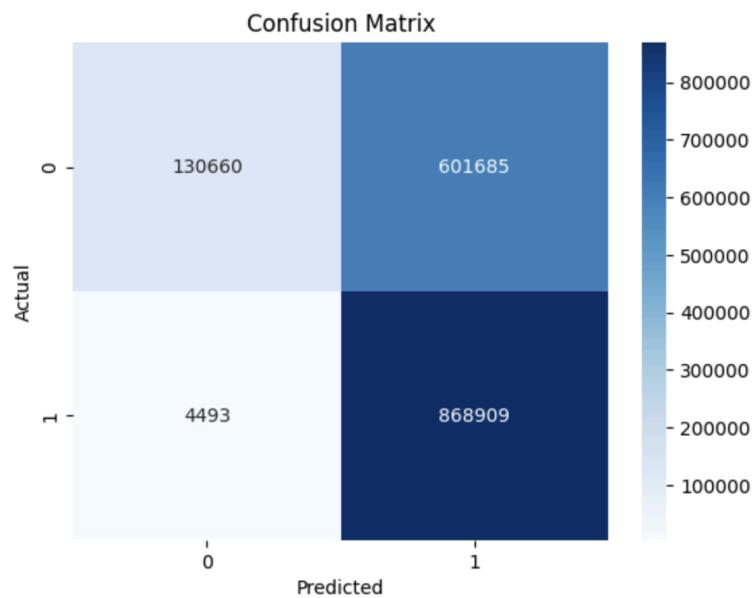
**Fig. 11:** Fragmento del historial de métricas durante el entrenamiento tales como el tiempo consumido por epoch, costo y exactitud del modelo en relación al dataset de entrenamiento y de validación.



**Fig. 12:** Diagrama de líneas que indica la evolución de la certeza del modelo en el tiempo, en base a los epochs. La línea azul sobre la naranja indica overfitting, y lo opuesto indica una capacidad pobre para generalizar.



**Fig. 13** La matriz de confusión que muestra la cantidad de positivos verdaderos en el cuadrante superior izquierdo, falsos negativos en el superior derecho, falsos positivos en el inferior izquierdo, y negativos verdaderos en el inferior derecho



**Fig. 14:** La matriz de confusión cuando el umbral se reduce a 0.05, causando que la precisión aumente hasta 97%.