Facultad
de
Ciencias

Mitigación del efecto de errores sistemáticos en clasificación de colisiones de partículas mediante aprendizaje automático

(Mitigating the Effect of Systematic Uncertainties in Particle Collision Classification using Machine Learning)

Trabajo de Fin de Grado
para acceder al

GRADO EN FÍSICA

Autor: Sergio Bolívar Gómez

Director: Francisco Matorras Weinig

Codirector: Pablo Martínez Ruiz del Árbol

Junio 2022

Agradecimientos

Como dice el refrán: «*es de bien nacido ser agradecido*». Por ello, me gustaría dedicar estas líneas a quienes me han acompañado en esta andadura.

En primer lugar, quisiera mostrar todo mi agradecimiento a mi director Francisco Matorras Weinig. Muchas gracias por tu incalculable ayuda, por tus consejos, por las innumerables charlas en tu despacho y por proponerme este trabajo. Me gustaría dar las gracias también a Pablo Martínez Ruiz del Árbol, mi codirector, por su inestimable ayuda con el código para que este trabajo llegase a buen puerto. Ha sido un placer trabajar con vosotros.

Siguiendo en el ámbito académico, me gustaría dar las gracias a los profesores de la Facultad de Ciencias por la formación recibida y a mis profesores del IES La Granja, sin los cuales posiblemente hoy no habría llegado hasta aquí.

En el ámbito personal, le tengo que estar eternamente agradecido a mis amigos. Por un lado, a mis amigos de "toda la vida". Gracias por confiar siempre en mí, entenderme y apoyarme en todo momento. Por otro lado, a los amigos que me ha regalado la carrera. Sin duda alguna, esta etapa no hubiese sido lo mismo sin vosotros. Espero que podamos compartir muchos más momentos juntos, ya sea en Santander, en Valencia, en Madrid, en Bruselas o allá donde nos depare el futuro.

Por último, y no por ello menos importante, muchas gracias a mi pequeña gran familia. Gracias por ser mi apoyo fundamental y soportarme durante estos cinco años. En especial, gracias a mi hermano y a mis padres, mis verdaderos maestros. Gracias por enseñarme a luchar y a no rendirme. Todo lo que tengo y lo que soy es gracias a vosotros.

A todos, muchísimas gracias.

Resumen

Los métodos multivariantes son un gran aliado para llevar a cabo tareas de clasificación cuando la dimensión del conjunto de datos es elevada. En particular, estas técnicas son muy utilizadas en física de partículas cuando se desea extraer una señal de interés de un fondo que es mucho más dominante. En general, estos métodos multivariantes ignoran el efecto de los errores sistemáticos en la fase de entrenamiento, resultando en modelos que quizá no se ajusten tanto a la realidad como se piensa. En este trabajo de fin de grado se exploran diversos métodos para incluir los efectos de los errores sistemáticos en la fase de entrenamiento de estos algoritmos. En particular, se propone un método reducido de *data augmentation*. Se ha encontrado que este método proporciona resultados bastante aceptables en ejemplos sintéticos, razón por la cual se ha comprobado su efectividad en un estudio real de búsqueda de materia oscura. En concreto, se ha aplicado para mitigar los errores sistemáticos que afectan a la energía de escala de los jets y a la energía transversa faltante. Se ha obtenido que, en algunos puntos de interés, el método propuesto es capaz de reducir más del 70 % el efecto de los errores sistemáticos. Más aún, se ha comprobado que para ciertos puntos de eficiencia permite disminuir la proporción de eventos de fondo mal clasificados como señal hasta en un 4 %, lo que supone un gran avance teniendo en cuenta que en el caso realista estudiado la señal queda enmascarada por un fondo varios órdenes de magnitud superior.

Palabras clave: aprendizaje automático, redes neuronales, errores sistemáticos, data augmentation, materia oscura.

Abstract

Multivariate methods are a great ally in performing classification tasks when the dimension of the data set is high. In particular, these techniques are widely used in particle physics when one wants to extract a signal of interest from a much more dominant background. In general, these multivariate methods ignore the effects of systematic uncertainties in the training phase, which means that the resulting models will not fit reality as well as one might expect. In this final project, several methods are explored to account for the effects of systematic uncertainties in the training phase of these algorithms. Among them, a reduced data augmentation method is proposed. This method has been shown to give quite acceptable results for synthetic examples, so its effectiveness has been tested in a real dark matter search study. In particular, it has been applied to reduce the systematic uncertainties in the jet energy scale and in the missing transverse energy. It has been shown that the proposed method can reduce the effects of the systematic uncertainties by more than 70 % at some points of interest. In addition, it has been found that in some other working points it can reduce the fraction of background events misclassified as signal by up to 4 %, which is a great advance considering that in the realistic case studied the signal is masked by a background several orders of magnitude larger.

Keywords: machine learning, neural networks, systematic uncertainties, data augmentation, dark matter.

Índice general

Capítulo 1

Introducción

When statistical errors dominated, systematic errors didn't matter much. In the days of particle factories and big data samples, they do.

Roger Barlow

La fiabilidad de los resultados en física de partículas experimental está fuertemente condicionada por la magnitud de las incertidumbres. Independientemente de si se mide la masa del bosón de Higgs, el factor g del muón o la masa del bosón W, es el tamaño de la incertidumbre lo que determina cómo de bueno es el resultado experimental y lo que permite hacer comparaciones entre resultados obtenidos por diferentes experimentos.

En el pasado, la precisión de los análisis estaba limitada por la falta de datos experimentales. En efecto, el hecho de no disponer de tecnología y experimentos suficientemente potentes como para generar enormes conjuntos de datos, hacía que la estadística disponible estuviese limitada. Esto provocaba un predominio de los errores estadísticos en los análisis, dejando en un segundo plano los errores sistemáticos.

En la actualidad, se puede decir que la situación es prácticamente la opuesta. La producción de datos experimentales no supone un gran problema, pues se dispone de experimentos que son capaces de generar cantidades ingentes de datos para su posterior análisis. Un ejemplo es el *Large Hadron Collider* (LHC) del *Conseil Européen pour la Recherche Nucléaire* (CERN), que es capaz de producir y registrar cientos de millones de colisiones de partículas por segundo. Efectivamente, que se pueda contar con más datos para llevar a cabo los análisis contribuye a la reducción de la incertidumbre estadística de los resultados, pero no permite abordar el problema de la incertidumbre sistemática [1]. En consecuencia, los errores sistemáticos son el factor limitante en muchos de los análisis reales, llegando incluso a ser los responsables de hacer fracasar un experimento. Sin ir más lejos, la confirmación o refutación de muchas teorías, como la del Modelo Estándar (SM, de *Standard Model*), depende de la precisión con que los datos experimentales se adecúan a las predicciones teóricas. Por ello, la mitigación del efecto de los errores sistemáticos es crucial.

En este sentido, los métodos multivariantes se postulan como un gran aliado para tratar el problema de los errores sistemáticos. En efecto, en el campo de la física de altas energías (HEP, de *High Energy Physics*) está muy extendido el uso de técnicas de aprendizaje automático (ML,

de *Machine Learning*) para, por ejemplo, la identificación de partículas [2, 3] o la búsqueda de nueva física [4, 5]. Estas técnicas más modernas proporcionan una mejora significativa con respecto a los análisis tradicionales basados en cortes, pues permiten manejar conjuntos de datos multidimensionales mucho más grandes y complejos, a la vez que extraen mucha más información de ellos a través de la búsqueda de patrones ocultos y correlaciones entre variables.

Sin embargo, a pesar de su importancia, los errores sistemáticos no suelen tenerse en cuenta a la hora de entrenar estos algoritmos. En su lugar, los modelos se entrenan con muestras de Monte Carlo (MC) ideales y los errores sistemáticos se incorporan a posteriori para estimar la incertidumbre que se propaga al resultado final. Este procedimiento aumenta la probabilidad de que los patrones de separación que se aprenden en el entrenamiento corran el riesgo de estar mal modelados, lo que resulta en una mayor sensibilidad a los errores sistemáticos [6]. Por ello, incorporar estos últimos directamente en la fase de entrenamiento de estos algoritmos es fundamental para minimizar la incertidumbre total [7].

Hasta la fecha, ha habido numerosos intentos para abordar este problema [6], pero no se ha logrado desarrollar un algoritmo que incluya el efecto de los errores sistemáticos en el entrenamiento y sea capaz de tener un buen rendimiento en una casuística general [8]. No obstante, se han diseñado algunas técnicas que han resultado ser de gran de utilidad para la mitigación del efecto de los errores sistemáticos en algunos casos concretos. Habitualmente, estas propuestas se dividen en dos grupos: las que se basan en construir algoritmos que sean completamente invariantes e insensibles al efecto de los errores sistemáticos, y las que proponen construir algoritmos que sean plenamente conscientes de la incertidumbre sistemática que afecta a los datos con el objetivo de ser menos sensible a sus efectos.

En el primero de los grupos, destacan las redes neuronales generativas adversarias [9, 10]. Esta técnica se basa en utilizar un sistema de dos redes neuronales confrontadas (discriminador y generador), que permite modificar la función de pérdidas durante el entrenamiento con el objetivo de que la distribución de la salida de la red sea robusta ante el efecto de los errores sistemáticos. Sin embargo, se ha visto que en algunos casos este método no presenta una ventaja significativa con respecto a las redes neuronales tradicionales, como puede comprobarse en el análisis efectuado en [11]. Otras propuestas que se integran dentro de este grupo son aquellas cuyas funciones de pérdida están diseñadas para maximizar directamente la significación estadística de la señal [12]. Ahora bien, uno de los inconvenientes de las técnicas anteriores es que no está muy claro como tratar el efecto combinado de varios errores sistemáticos [8].

Con el propósito de dar respuesta al problema anterior, se ha propuesto recientemente INFERNO (de *Inference-Aware Neural Network*) [13]. Esta técnica aborda el problema de los errores sistemáticos de una manera innovadora, pues trata de minimizar directamente la incertidumbre de los parámetros de interés (obtenidos a partir de la función de verosimilitud) mediante modificaciones en la función de pérdidas [8, 14]. Se ha demostrado que esta metodología tiene muy buen rendimiento en problemas sintéticos, anulando completamente el efecto combinado de múltiples incertidumbres sistemáticas [14]. Sin embargo, su eficacia aún no ha sido comprobada en casos realistas [13].

En el segundo de los grupos, se encuentran principalmente las técnicas de *data augmentation*. Éstas consisten en aumentar el conjunto de entrenamiento mediante la inclusión de observaciones que se vean afectadas por los efectos de los errores sistemáticos. Así, se consigue que el algoritmo se exponga a datos con mayor variabilidad durante la fase de entrenamiento, de manera que a la hora de hacer nuevas predicciones el modelo sea menos sensible a los efectos sistemáticos.

1.1. Objetivos. Estructura del TFG

El objetivo principal de este Trabajo Fin de Grado (TFG) es triple: (1) proponer algunos métodos que permitan mitigar el efecto de los errores sistemáticos en la fase de entrenamiento de los métodos multivariantes, (2) probar su eficacia en ejemplos sintéticos, y (3) comprobar la efectividad de algunas de estas propuestas en un caso realista de un análisis de colisiones del acelerador LHC.

La estructura de la memoria es la siguiente. En el Capítulo 2 se describen algunos aspectos teóricos que fundamentan el trabajo y a los que se hará alusión de manera frecuente en él. En el Capítulo 3 se proponen una serie de métodos para la mitigación del efecto de los errores sistemáticos y se someten a prueba en algunos ejemplos sintéticos. El Apéndice A contiene algunos resultados de estas pruebas. En el Capítulo 4 se aplican algunos de los métodos propuestos anteriormente en un caso realista de búsqueda de materia oscura con datos del LHC. Finalmente, en el Capítulo 5 se incluyen las conclusiones.

1.2. Problema físico: búsqueda de materia oscura en CMS

El caso realista analizado en este TFG tiene que ver con la búsqueda de Materia Oscura (DM, de *Dark Matter*) en asociación con quarks top a partir de las colisiones protón-protón que tienen lugar en el experimento CMS (*Compact Muon Solenoid*) del LHC, en el CERN.

1.2.1. Experimento CMS del LHC

El LHC (*Large Hadron Collider*) [15] es el mayor acelerador de partículas del mundo. Se ubica en el CERN, en la frontera entre Francia y Suiza, y está constituido por un anillo subterráneo de 27 km de circunferencia construido sobre lo túneles del antiguo LEP (*Large Electron-Positron Collider*). En él, se aceleran haces de protones hasta energías relativistas (∼13.6 TeV) y se hacen colisionar en cuatro puntos estratégicos, en cada uno de los cuales se sitúa un experimento: LHCb, ATLAS, ALICE y CMS.

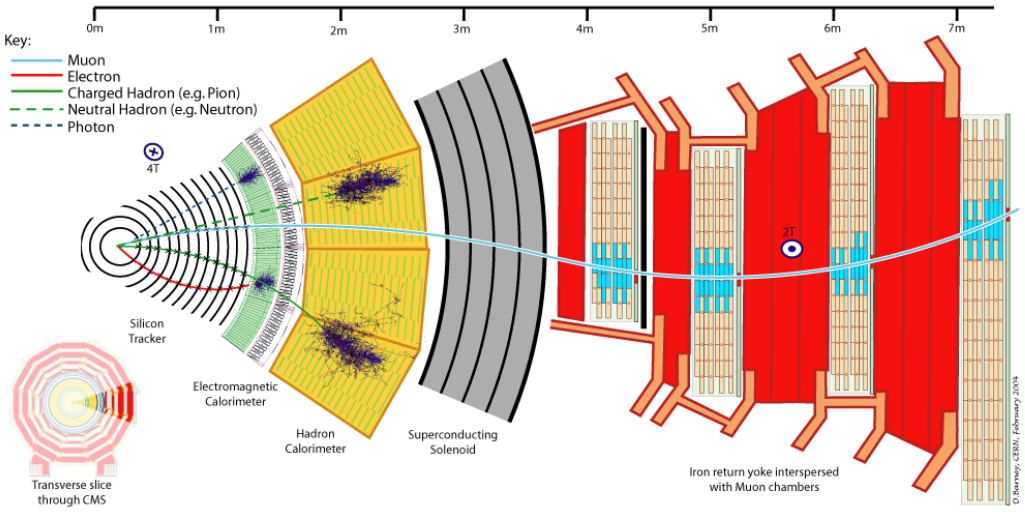

Figura 1.1: Estructura interna del detector CMS [16]. Se muestra la sección transversal del detector, así como los subdetectores que lo integran.

En particular, el experimento CMS [17] está formado por un imán solenoidal que permite curvar las trayectorias de las partículas, y cuatro subdetectores principales distribuidos en capas, a saber: el detector de trazas, el calorímetro electromagnético, el calorímetro hadrónico y el sistema de muones (véase la Figura 1.1). El detector de trazas o *tracker* se encarga de reconstruir las trazas de todas las partículas cargadas que alcanzan el detector. El calorímetro electromagnético permite medir la energía de los electrones y los fotones (que depositan toda su energía en este subdetector), mientras que el calorímetro hadrónico es el encargado de medir la energía de los hadrones producidos en la colisión. Finalmente, el sistema de muones permite detectar los muones y medir su momento.

La geometría del detector CMS es, por tanto, cilíndrica, coincidiendo el eje longitudinal (eje Z) con la dirección en la que se hacen colisionar los protones. El eje X se escoge en la dirección que apunta hacia el centro del anillo del LHC, mientras que el eje Y se toma apuntando hacia arriba. El ángulo azimutal ϕ es el ángulo entre el eje X y el eje Y, mientras que el ángulo polar θ es el ángulo entre los ejes Z e Y. En lugar de utilizar el ángulo polar, se emplea la pseudorapidez, que es una cantidad invariante Lorentz definida como $\eta = -\ln(\tan(\theta/2))$.

Son numerosos los observables físicos que se pueden medir gracias a este detector. Entre ellos, el más importante para nuestro análisis es el momento transverso, p_T, que no es más que la componente del momento lineal en el plano transversal del detector (perpendicular al eje Z).

Cabe señalar que existen partículas eléctricamente neutras que atraviesan CMS y son indetectables, como es el caso de los neutrinos o las hipotéticas partículas de DM. No obstante, la energía de estas partículas que atraviesan los detectores de CMS sin dejar traza puede estimarse gracias al principio de conservación del momento. En efecto, en el LHC los haces de protones se hacen colisionar a lo largo del eje longitudinal de CMS, de manera que la suma vectorial de los momentos transversos después de la colisión ha de ser nula. Si esta suma no es nula, entonces es un indicador de que se han producido partículas y no han sido registradas por el detector. Precisamente, se define el momento transverso faltante como

$$\overrightarrow{\not{p}_T} = -\sum_{\substack{\text{particulas}\\ \text{visibles}}} \overrightarrow{p_T}.$$

La energía transversa faltante (MET, de *Missing Transverse Energy*) se define como el módulo del momento transverso faltante y representa la energía que no se detecta en el detector, pero que se espera debido a la conservación del momento.

1.2.2. Materia Oscura en CMS

La existencia de materia oscura puede inferirse, por ejemplo, a partir de las anomalías observadas en las curvas de rotación de las galaxias [18] o las anisotropías en la temperatura del fondo cósmico de microondas [19]. Sin embargo, hasta la fecha, ha sido imposible detectarla de manera experimental. Por ello, el análisis de procesos físicos donde se piensa que se podría producir este tipo de materia es uno de los grandes retos de la comunidad científica en la actualidad. En este trabajo, se aplican algunos de los métodos de mitigación del efecto de los errores sistemáticos a datos enmarcados en la búsqueda de materia oscura a través de colisiones de partículas del Modelo Estándar en el experimento CMS del LHC, un tipo de búsqueda llamada de "colisionador", que difiere de las búsquedas directas [20] e indirectas [21].

En muchos de los modelos de nueva física (BSM, de *Beyond Standard Model*), la materia oscura es una partícula estable, no bariónica, muy masiva y que interacciona débilmente, pero no

electromagnéticamente. Este tipo de partículas se conocen como WIMPs (*weakly-interacting massive particles*) [22]. El hecho de ser eléctricamente neutras, hace que las partículas de materia oscura sean indetectables en el experimento CMS. En consecuencia, su presencia ha de inferirse mediante desbalances de momento en el plano transverso del detector. El principal problema es que a este desbalance del momento transverso contribuyen también los neutrinos procedentes de otros procesos físicos del Modelo Estándar, lo que dificulta enormemente la tarea de aislar la hipotética señal de materia oscura (en caso de manifestarse de esta manera).

Un opción para restringir las regiones de búsqueda en el espacio de fases y disminuir el efecto de los procesos de fondo es la aplicación de cortes en algunas de las variables involucradas en el análisis. No obstante, estos cortes muchas veces son insuficientes y es necesario recurrir a métodos más complejos de análisis multivariante. Aún así, la tarea de discriminación entre señal y fondo sigue siendo dificultosa, ya que la variedad de procesos de fondo es muy amplia.

Por ello, es habitual buscar la materia oscura en asociación con otras partículas. Con esta técnica, se restringen significativamente los sucesos analizados, lo que permite reducir notablemente los procesos físicos de fondo (aunque no los elimina por completo).

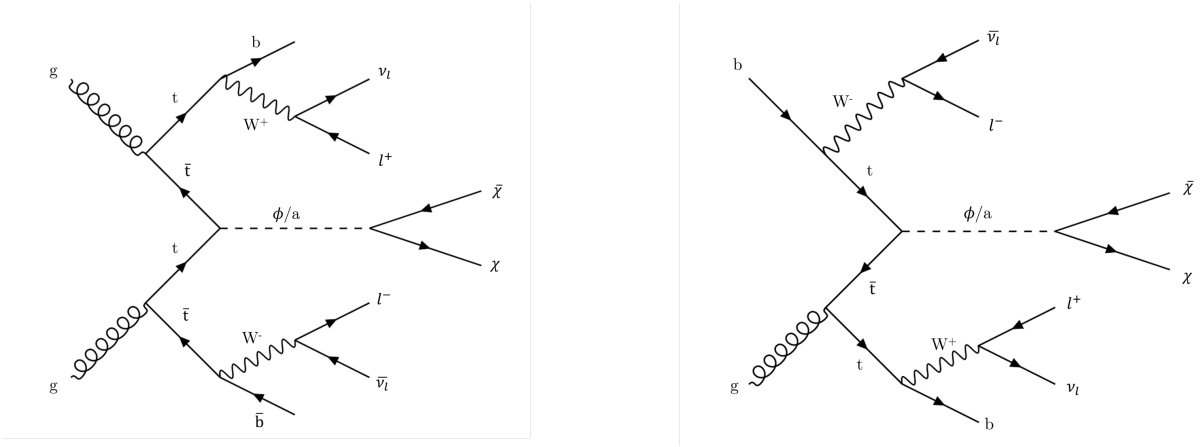

Figura 1.2: Diagramas de Feynman típicos de un proceso $t\bar{t}$+DM (izquierda) y $t/\bar{t}$+DM (derecha).

En particular, en este trabajo se analizan datos procedentes de la búsqueda de materia oscura en asociación con quarks top en el estado final dileptónico. La razón es que el quark top, al ser el más masivo de los quarks, es el que tiene una interacción más intensa con el bosón de Higgs. Por otro lado, las partículas de materia oscura adquirirían su masa por el mecanismo de Higgs e interaccionarían intensamente con él porque son muy masivas. Por tanto, cabe esperar que exista una relación "especial" entre los quarks top y las partículas de materia oscura.

Fundamentalmente, se manejan datos del canal de producción de materia oscura en asociación con un único quark top ($t/\bar{t}$+DM), aunque también en asociación con dos quarks top ($t\bar{t}$+DM). Los diagramas de Feynman típicos para estos procesos se encuentran representados en la Figura 1.2. Se ha escogido el estado final dileptónico porque, aunque sea menos favorable desde el punto de vista estadístico, es un canal mucho más limpio de procesos de fondo.

Capítulo 2

Marco teórico

En este capítulo se presentan los conceptos teóricos en los que se fundamenta principalmente el trabajo. En primer lugar, se hace una introducción a los métodos de análisis multivariante y se describe la manera de evaluar su rendimiento. A continuación, se proporciona un criterio para distinguir entre errores estadísticos y sistemáticos. Finalmente, se explican los errores sistemáticos involucrados en el caso realista analizado en esta memoria.

2.1. Métodos de Análisis Multivariante

El análisis multivariante (MVA, de *Multivariate Analysis*) es una herramienta muy utilizada en el campo de la física de partículas [23, 24]. En efecto, una de las tareas más habituales a la hora de analizar los datos procedentes de colisiones de partículas consiste en extraer la señal de interés de un fondo mucho más dominante. En ocasiones, hay una única variable que es suficientemente poderosa como para poder discriminar la señal del fondo. Un ejemplo de esta situación es el decaimiento de un bosón de Higgs a dos fotones, donde la masa invariante del par de fotones es la variable discriminante [25, 26]. Sin embargo, lo más habitual es que sea necesario combinar el poder discriminante de varias variables para lograr la adecuada extracción de la señal, como ocurre en el análisis efectuado en este trabajo. Es en esta última situación donde entran en juego los métodos de clasificación multivariantes propios del análisis discriminante.

El marco de trabajo para las técnicas de clasificación supervisada es muy simple. Se dispone de un conjunto de observaciones, representadas cada una de ellas por una serie de variables numéricas, de las que se conoce la clase a la que pertenecen. En nuestro caso particular, esta muestra la conforman observaciones procedentes de simulaciones de MC de los procesos de fondo y de señal más relevantes. Por construcción, cada uno de estos eventos está etiquetado con su correspondiente categoría, a saber: señal $t/\bar{t}$+DM, señal $t\bar{t}$+DM o fondo. Este conjunto inicial de datos se puede dividir en tres subconjuntos disjuntos. Por un lado, se dispone de un conjunto de entrenamiento (*training sample*), que se utiliza para ajustar los parámetros del modelo. Por otro lado, se dispone de un conjunto de prueba (*test sample*), que permite evaluar el rendimiento del modelo entrenado y detectar posibles casos de sobreentrenamiento[1]. Finalmente, se puede extraer un conjunto de validación (*validation sample*), que permitirá comparar diferentes modelos entre sí.

[1] ocurre cuando el clasificador aprende a separar correctamente las observaciones de la muestra de entrenamiento, pero no la generalidad.

Una vez entrenado un modelo aceptable, la utilidad de estas técnicas es que pueden inferir de manera autónoma la categoría a la que pertenecen datos independientes, como por ejemplo los datos reales colectados por el detector CMS del LHC.

Actualmente, existen numerosos paquetes y librerías que implementan todo tipo de métodos para el análisis de datos multidimensionales. En particular, en este trabajo se han manejado los siguientes:

- **TMVA** (*Toolkit for Multivariate Data Analysis*) [27], que es una librería desarrollada para ROOT que implementa una multitud de métodos de análisis multivariante. Está orientada al tratamiento de datos propios de la física de altas energías, y especialmente útil para la resolución de problemas de clasificación y regresión. Al igual que ROOT, TMVA está desarrollado en C++, aunque posee una implementación en Python, que es la que se ha utilizado en este trabajo para obtener los resultados expuestos en el Capítulo 4.
- **KERAS** [28], que es una librería que proporciona una interfaz de RStudio y de Python para implementar redes neuronales artificiales. La versión de RStudio se ha utilizado en las fases de prueba de los métodos propuestos en el Capítulo 3. La versión de Python se ha utilizado en combinación con TMVA para efectuar el análisis realista.

A continuación, se describen los dos métodos de aprendizaje supervisado empleados en este análisis: las redes neuronales artificiales (ANNs, de *Artificial Neural Networks*) y los árboles de decisión "potenciados" (BDTs, de *Boosted Decision Trees*).

2.1.1. Redes Neuronales Artificiales

Las redes neuronales artificiales (ANNs) son un modelo computacional que está inspirado en el mecanismo de aprendizaje seguido por el cerebro de los seres vivos [29, 30]. En particular, son útiles en problemas no lineales consistentes en la búsqueda de patrones y agrupaciones dentro de conjuntos de datos multidimensionales.

En analogía con la estructura del cerebro humano, las redes neuronales artificiales están compuestas por una serie de nodos o unidades computacionales interconectadas, llamadas **neuronas**, que simulan las neuronas propiamente biológicas. Las conexiones entre estos nodos se basan en una **topología de red** predefinida (encargada de regular la transmisión de la información) y llevan asociado un **peso**, que refleja la importancia de dicha conexión dentro de la red [31].

En definitiva, las neuronas no son más que unidades computacionales que reciben una señal de entrada (*inputs*) y que calculan la señal de salida (*output*) mediante una función no lineal, conocida como **función de activación**, aplicada a la suma ponderada de los *inputs*. Esta ponderación se hace en base a los pesos asociados a cada conexión.

Existen numerosos tipos de redes neuronales, entre las que caben destacar las redes neuronales convolucionales (CNN), las recurrentes (RNN), las de base radial (RBF) y las perceptrón multicapa (MLP). Son estas últimas las que serán objeto de estudio en este trabajo.

En las redes neuronales perceptrón multicapa (MLP, de *Multilayer Perceptron*) [31] las neuronas se agrupan en varias **capas**, entre las que cabe destacar la capa de entrada (dedicada a la adquisición de los datos) y la capa de salida (destinada a devolver los resultados). Las capas que no son de entrada ni de salida reciben el nombre de capas intermedias u ocultas. El aspecto más importante de este tipo de redes neuronales es su arquitectura, pues se asume que todos los nodos de una capa concreta están conectados con los de la siguiente capa. Así, cada nodo

de la capa de entrada alimenta la red hacia delante (*feed-forward network*), de manera que cada nodo en las capas ocultas recibe gradualmente una entrada procedente de los nodos de las capas inmediatamente previas y computa un valor de salida para cada nodo de la siguiente capa. El proceso termina cuando se alcanza la capa de salida, cuyo *output* es el resultado de una transformación progresiva de los valores de entrada. La Figura 2.1 muestra un ejemplo de redes neuronales tipo MLP, con cuatro neuronas en cada una de sus tres capas ocultas.

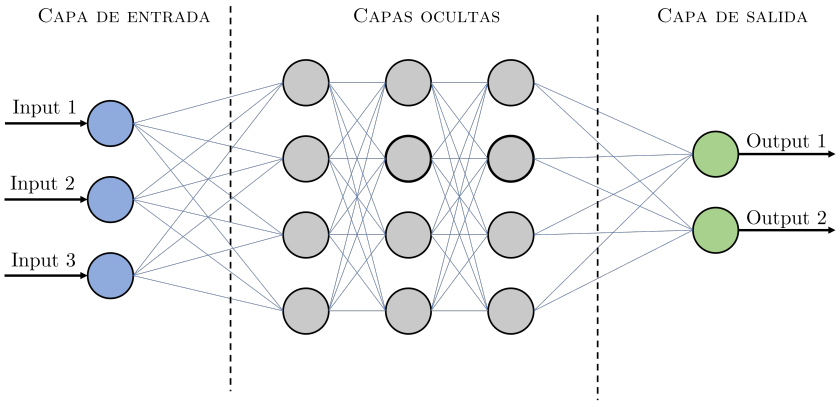

Figura 2.1: Perceptrón multicapa con topología *3:4:4:4:2*, es decir, 3 neuronas en la capa de entrada, 4 neuronas en cada una de las tres capas ocultas, y 2 neuronas en la capa de salida.

Supongamos una red multicapa que posee una capa oculta arbitraria k con n neuronas, todas ellas con la misma función de activación f_k. Asumamos que existe una capa oculta inmediatamente anterior $k-1$ con m neuronas. Formalmente, cada neurona j, con $j \in \{1,\ldots,n\}$, de la capa oculta k recibirá un *input* x_i, con $i \in \{1,\ldots,m\}$, de cada una de las m neuronas de la capa $k-1$ y los combinará linealmente multiplicando cada uno de ellos por el correspondiente peso de la conexión $\omega_{i,j}$[2]. A continuación, cada neurona j aplica a la combinación lineal anterior una función de activación no-lineal, tal y como se muestra en la Eq.(2.1). El resultado y_j de la composición anterior es el *output* de la neurona j-ésima, que servirá para alimentar la siguiente capa oculta. Este proceso de repite sucesivamente hasta alcanzar la capa de salida.

$$y_j = f_k\left(\sum_{i=1}^{m} \omega_{j,i} x_i\right) \tag{2.1}$$

Se ha comentado previamente que lo que caracteriza a las redes neuronales artificiales es su capacidad para aprender. Al igual que los estímulos externos son necesarios para el aprendizaje humano, el estímulo externo en las ANNs lo proporciona el conjunto de entrenamiento, que contiene ejemplos de *inputs* para los que el *output* deseado es conocido. En este sentido, el aprendizaje en las ANNs se produce al actualizar los pesos que conectan las neuronas. Por tanto, el objetivo del proceso de entrenamiento consiste en encontrar los pesos de cada una de las conexiones de la red, representado por el vector $\boldsymbol{\omega}$, de manera que el *output* obtenido sea lo más cercano posible al deseado. La manera de comprobar cómo de óptimos son los pesos obtenidos es a través de la función de pérdidas, que típicamente no es más que el error cuadrático

[2]el peso $\omega_{j,i}$ refleja la fuerza de la conexión de la neurona i-ésima de la capa $k-1$ con la neurona j-ésima de la capa k.

medio (MSE, de *Mean Squared Error*)[3]. Matemáticamente, si el conjunto de entrenamiento tiene N ejemplos, se trata de resolver el siguiente problema de minimización

$$\underset{\boldsymbol{\omega}\in\mathbb{R}^p}{\text{Minimizar}} \quad \mathrm{E}(\boldsymbol{\omega}) = \frac{1}{N}\sum_{i=1}^{N}\left(\theta_i(\boldsymbol{\omega}) - \hat{\theta}_i\right)^2 \tag{2.2}$$

donde p es el número de conexiones totales en la red, $\theta_i(\boldsymbol{\omega})$ es la predicción del modelo para el evento i-ésimo del conjunto de entrenamiento y $\hat{\theta}_i$ es su valor real.

Para resolver el problema anterior, se sigue un proceso iterativo en el que se van actualizando los pesos de la red de manera que se reduzca la función de pérdidas. La expresión general para el vector de pesos en la iteración $(k+1)$-ésima es:

$$\boldsymbol{\omega}^{(k+1)} = \boldsymbol{\omega}^{(k)} + \rho_k d^{(k)}, \tag{2.3}$$

donde $d^{(k)}$ es un vector que representa una dirección de descenso (porque cumple que a lo largo de esta dirección la función E decrece, al menos para desplazamientos suficientemente pequeños) y ρ_k es un número real que indica la longitud del desplazamiento en esa dirección.

Existen varias técnicas que permiten actualizar los pesos de la red, entre las que caben destacar las siguientes:

- **Descenso de gradiente:** Esta técnica de aprendizaje consiste en calcular la dirección de descenso de manera que sea proporcional al gradiente de la función de pérdidas con respecto a los pesos de la red [32, 33]. En concreto, la actualización de los pesos viene dada por
$$\Delta\boldsymbol{\omega}^{(k)} := \boldsymbol{\omega}^{(k+1)} - \boldsymbol{\omega}^{(k)} = -\eta\nabla_{\boldsymbol{\omega}}E(\boldsymbol{\omega}^{(k)}),$$
donde se puede se puede hacer la identificación $\rho_k \equiv \eta$ y $d^{(k)} = -\nabla_{\boldsymbol{\omega}}E(\boldsymbol{\omega}^{(k)})$ en la Ec.(2.3). La constante positiva η recibe el nombre de tasa de aprendizaje o *learning rate*.

- **Descenso de gradiente con momento:** Esta técnica se desarrolló para aumentar la tasa de convergencia del método anterior [34]. En particular, consiste en que la actualización del vector de pesos en una iteración dada sea una combinación lineal del gradiente de la función de pérdidas y la modificación de pesos correspondiente a la iteración anterior [33]. Añadiendo este término adicional, se da cierta inercia al método y se evita el estancamiento en mínimos locales. Formalmente,
$$\Delta\boldsymbol{\omega}^{(k)} = -\eta\nabla_{\boldsymbol{\omega}}E(\boldsymbol{\omega}^{(k)}) + \beta\Delta\boldsymbol{\omega}^{(k-1)},$$
donde se puede se puede hacer la identificación $\rho_k \equiv 1$ y $d^{(k)} = -\nabla_{\boldsymbol{\omega}}E(\boldsymbol{\omega}^{(k)}) + \beta\Delta\boldsymbol{\omega}^{(k-1)}$ en la Ec.(2.3). La constante positiva β representa el momento.

- **Descenso de gradiente estocástico:** La técnica es similar al descenso de gradiente, con la salvedad de que el gradiente no se calcula con respecto a todo el conjunto de datos de entrenamiento, sino que se hace sobre un subconjunto aleatorio de los mismos [35]. Esto consigue reducir drásticamente la complejidad del método y lograr una convergencia mucho más rápida.

[3] salvo que se indique lo contrario, en todo lo que sigue consideraremos que la función de pérdidas es el MSE.

Como puede verse, los métodos difieren esencialmente en la manera de calcular la dirección de descenso y el paso. No obstante, la mayoría de los algoritmos de aprendizaje siguen alguna de las técnicas descritas anteriormente o alguna ligera modificación de las mismas. En particular, en este trabajo se ha utilizado el algoritmo *Adam* [36], que sigue un método de descenso de gradiente estocástico combinado con momento.

Un aspecto importante, quizá pasado por alto, son las propiedades que deben cumplir las funciones de activación de las neuronas. Normalmente, la función de activación escogida depende del problema que se quiere resolver. Así, en problemas de clasificación que no son linealmente separables es habitual recurrir a funciones de activación no lineales. Además, para garantizar el uso de técnicas de descenso del gradiente, se requiere que estas funciones sean continuamente diferenciables. Entre las funciones de activación más destacadas se encuentran las funciones sigmoide, las tangentes hiperbólicas, las rectificador (*ReLu*) o las funciones *SoftMax* [27].

En base a todo lo anterior, son numerosos los hiperparámetros que hay que fijar y/o ajustar a la hora de utilizar las redes neuronales para abordar un problema de clasificación. Entre ellos, caben destacar: el número de neuronas en la capa de entrada (generalmente, coincide con el número de variables involucradas) y en la capa de salida (habitualmente, una por cada categoría posible), el número de capas ocultas y las neuronas en cada una de ellas, el tipo de función de activación para cada una de las capas, la función de error, la tasa de aprendizaje, el momento... Los hiperparámetros de la redes neuronales artificiales utilizadas en el análisis efectuado en este trabajo se detallan en el Capítulo 3 (para el caso sintético) y en el Capítulo 4 (para el caso real).

2.1.2. Boosted Decision Trees

Los árboles de decisión "potenciados" (BDTs), como su propio nombre indica, están basados en árboles de decisión (DTs, de *Decision Trees*). Los DTs son un tipo de clasificador que considera el conjunto de entrenamiento y lo divide de manera recursiva en base a las variables que lo definen. En cuanto a su estructura, los DTs están formados por varios nodos, en los que se produce la división de los datos, y hojas, que son nodos terminales en los que ya no se producen más divisiones. La división del conjunto de datos en cada nodo se efectúa en base a un cierto criterio, como puede ser minimizar la entropía de Shannon o maximizar la ganancia de información [27]. El proceso recursivo de división de los datos del conjunto de entrenamiento se realiza hasta que se satisfacen algunas condiciones de parada, o hasta que todas las observaciones en cada hoja pertenezcan a la misma categoría. El procedimiento anterior define una serie de reglas de decisión basadas en el conjunto de entrenamiento que permiten clasificar nuevas observaciones.

Los BDTs se basan en agregar varios DTs (proceso conocido como *boosting*), lo que permite construir un clasificador más potente. En particular, la salida del BDT es la suma ponderada de las salidas de los DTs que lo componen, siendo el peso que se le asigna a cada DT dependiente de la tasa de éxito que este clasificador obtenga [27].

2.1.3. Evaluación del rendimiento

Una vez que se ha entrenado un modelo predictivo, las muestras de test son útiles para evaluar su rendimiento o detectar posibles casos de sobreentrenamiento. En este contexto, surge la siguiente pregunta: ¿qué métricas debemos analizar para concluir si el clasificador es aceptable? La respuesta no es única, y dependerá principalmente del problema y de las características de los datos con los que se esté trabajando.

En el caso de los experimentos diseñados para la búsqueda de nueva física, se pretende maximizar la significación estadística de la muestra de la señal sobre la muestra del fondo, en lugar de obtener la mejor precisión de clasificación. En otras palabras, se busca maximizar el número de eventos de señal clasificados correctamente, a la vez que se minimizan los eventos de fondo clasificados incorrectamente. En consecuencia, la tasa de aciertos en la clasificación (*accuracy*, del inglés) no es una métrica muy representativa en este tipo de problemas. La razón es que condensa en un único número el rendimiento global del modelo, cuando realmente hay interés en conocer el rendimiento en una región muy concreta.

Una manera de evaluar el rendimiento de un modelo de forma más localizada es a través de la *Receiver Operating Curve* (curva ROC). Aunque esta curva admite varias representaciones, en este trabajo seguiremos el criterio expuesto en la Figura 2.2. Por un lado, en el eje vertical se representa la especificidad, esto es, la tasa de verdaderos negativos (los eventos de fondo que se clasifican correctamente como fondo). Por otro lado, en el eje horizontal se representa la sensibilidad, que se corresponde con la tasa de verdaderos positivos (los eventos de señal clasificados correctamente). En el ámbito de la física de altas energías, la especificidad se corresponde con el nivel de rechazo de fondo, mientras que la sensibilidad refleja la eficiencia de la señal.

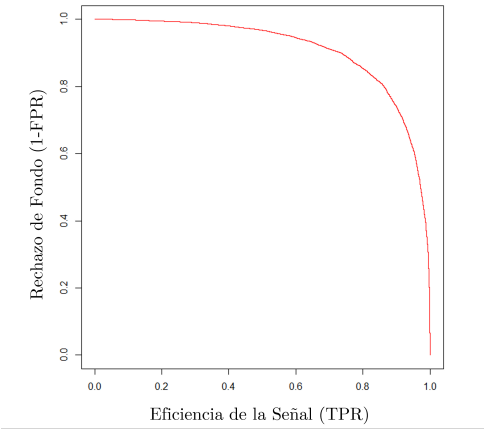

Figura 2.2: Ejemplo de curva ROC. Por un lado, en el eje vertical se representa el nivel de rechazo de fondo, también llamado especificidad o TNR (*True Negative Rate*, del inglés). Equivalentemente, podemos referirnos a él como 1-FPR, donde FPR (*False Positive Rate*, del inglés) es la tasa del falsos positivos. Por otro lado, en el eje horizontal se representa la eficiencia de la señal, que se corresponde con la sensibilidad o TPR (*True Positive Rate*, del inglés).

En la línea de lo comentado anteriormente, la región de interés de la curva ROC en experimentos de búsqueda de nueva física es la superior. La razón es que esta región permite un alto rechazo de fondo. Idealmente, se estaría interesado en alcanzar la esquina superior derecha, ya que en este caso la significación estadística de la señal sería muy elevada.

En particular, en este trabajo se busca desarrollar un método que tenga en cuenta el efecto de los errores sistemáticos de manera que: (1) fijado un nivel de rechazo de fondo, se tenga una eficiencia de señal mejor que la obtenida con los métodos convencionales, (2) fijada la eficiencia de la señal, se tenga un nivel de rechazo de fondo superior al obtenido con métodos que no tienen en cuenta los efectos de los errores sistemáticos.

Finalmente, cabe mencionar que otra métrica muy utilizada para evaluar el rendimiento de un clasificador es el área bajo la curva ROC (AUC). Sin embargo, la AUC no es una buena métrica en nuestro caso porque condensa el rendimiento global del modelo, dando importancia a regiones que para nuestro análisis pueden no ser interesantes.

2.2. Errores sistemáticos y estadísticos

En general, cuando se realiza la medición de una magnitud física hay dos fuentes principales de incertidumbre: los errores estadísticos y los errores sistemáticos. En la práctica, esta dicotomía no está claramente definida [1], lo que en ocasiones da lugar a interpretaciones erróneas de estos conceptos.

Por un lado, los errores estadísticos tienen su origen en la estocasticidad inherente al proceso de medición. En efecto, las fluctuaciones estocásticas son las responsables de la obtención de resultados diferentes cuando una misma medida se realiza varias veces. La incertidumbre estadística es, por tanto, una medida del rango de estas variaciones. Por definición, las variaciones estadísticas entre las medidas del mismo fenómeno en igualdad de condiciones son incorreladas. En consecuencia, la repetición de una medida múltiples veces permite restringir la incertidumbre estadística. Algunas fuentes de error estadístico son la tasa de conteo en detectores o las variaciones aleatorias en el sistema que se esté observando [1].

Por otro lado, los errores sistemáticos tienen su origen en la naturaleza del aparato de medida, en las hipótesis realizadas por el observador e incluso en el modelo empleado para analizar los datos. A diferencia de los errores estadísticos, los errores sistemáticos pueden estar correlacionados de unas medidas a otras, de manera que repetir la medición no será de utilidad para reducir este tipo de incertidumbre. Algunas fuentes de errores sistemáticos son la calibración de los aparatos de medida, la resolución finita de los aparatos de medida, la aceptancia de los detectores o los propios parámetros del modelo utilizado para interpretar los datos [1].

En consecuencia, los errores sistemáticos requieren una atención especial ya que, aunque en ocasiones sean indetectables y/o inmensurables, pueden hacer fracasar el experimento. Idealmente, habría que hacer un análisis exhaustivo para analizar todas los posibles fuentes de error sistemático involucradas en el experimento para, después, tenerlas en cuenta de alguna manera en el análisis de los resultados.

Por todo ello, los errores sistemáticos constituyen un factor limitante a la hora de realizar un experimento, razón por la que buscar técnicas para mitigarlos es un reto en la actualidad.

2.2.1. Muestras afectadas por los errores sistemáticos

En esta sección se describe cómo se generan las muestras afectadas por los errores sistemáticos. Supongamos que tenemos una muestra nominal, y que estamos interesados en considerar el error sistemático que afecta a una de sus variables, llamémosla x. El procedimiento que se sigue para generar las muestras con la sistemática *up* y sistemática *down* para la variable x consiste fundamentalmente en transformar la variable x de cada uno de los sucesos de la muestra nominal en $x+\Delta x$ (sistemática *up*), y en $x-\Delta x$ (sistemática *down*), donde Δx es la distorsión debida al error sistemático. El valor de Δx puede ser constante, o puede depender del propio valor de x. Por ejemplo, podría ocurrir que el efecto del error sistemático venga dado por un $p\,\%$ del valor de x, por lo que en este caso se tendría que $\Delta x = (p/100)\cdot x$. Este caso es el que se considerará más habitualmente en los ejemplos sintéticos.

2.2.2. Efecto de algunos errores sistemáticos en un caso realista

En un análisis realista de física de partículas son numerosos los errores sistemáticos que entran en juego. Sin embargo, en el caso estudiado en este trabajo, interesan sobre todo aquellos que afectan a la escala de energía de los jets [37] y a la medida de la energía transversa faltante [38].

La escala de energía de los jets (JES, de *Jet Energy Scale*) es una de las magnitudes que permite parametrizar el rendimiento de los detectores de CMS [39]. En particular, la JES es un indicador del sesgo con que el detector mide los valores reales de energía de los jets. En otras palabras, es un factor de calibración que ha de aplicarse a las medidas reales de la energía de los jets para corregirlas por los efectos inherentes a su reconstrucción. La manera de cuantificar esta cantidad es haciendo una representación gráfica en la que se enfrenta la energía reconstruida de los jets frente a la energía simulada (real) de los mismos. En concreto, se espera una línea recta ($y = x$) si el detector está correctamente calibrado y el algoritmo de *clustering* funciona correctamente. Las variaciones con respecto a la situación ideal anterior se emplean para calcular la JES. Un aspecto importante es que la JES no es constante para un detector, sino que depende del momento transverso, p_T, de los jets y de la pseudorapidez, η.

Los errores sistemáticos que afectan a la JES son, por tanto, muy importantes, pues pueden alterar notablemente los resultados del experimento al propagarse en la medición de muchos parámetros. En particular, las variaciones en la JES se propagan a otras variables discriminantes para nuestro análisis, como la MET. La incertidumbre sistemática se espera que sea inferior al 3 % en el espacio de fases considerado en la mayoría de los análisis ($p_T > 30$ GeV, $|\eta| < 5.0$), e incluso inferior al 1 % en la región $|\eta| < 1.3$ para jets con $p_T > 30$ GeV [40]. Los errores sistemáticos que afectan a la JES tienen su origen, principalmente, en la calibración y en el algoritmo de reconstrucción.

Por otro lado, la MET es una variable bastante discriminante en el problema físico abordado en este trabajo. Esto se debe a que, aunque haya algunos procesos de fondo del SM que produzcan MET a través de la producción de neutrinos, la señal $t/\bar{t}+DM$ tendrá, además de la contribución a la MET de sus propios neutrinos, una contribución extra del par $\chi\overline{\chi}$ que se produce. Por ello, se espera que el espectro de la MET alcance valores más altos para la señal que para los procesos de fondo del SM.

En consecuencia, analizar el efecto de los errores sistemáticos en esta variable es crucial. Algunas fuentes de errores sistemáticos que afectan a la MET son, por ejemplo, el ruido en los calorímetros, el propio proceso de reconstrucción o la calibración.

Finalmente, cabe destacar que la generación de las muestras afectadas por la sistemática en un caso realista es más complicado que lo expresado en a Sección 2.2.1. La razón es que los efectos de los errores sistemáticos que afectan a ciertas variables, se propagan también a las variables derivadas. Más aún, los errores sistemáticos también pueden alterar el número de sucesos considerados en el análisis. En efecto, si se establece un corte en el momento transverso de los jets en $p_T > 20$ GeV, podría ocurrir que el efecto de los errores sistemáticos haga que jets que tenían originalmente $p_T = 21$ GeV pasen a tener $p_T = 19$ GeV y, por tanto, no se consideren en el análisis. Lo mismo podría ocurrir al revés, esto es, que sucesos que antes no se consideraban en el análisis, pasen a tenerse en cuenta.

En este sentido, hay grupos de trabajo especializados en el experimento CMS que realizan estas tareas y proporcionan tablas de corrección e incertidumbres de estos parámetros para que cada análisis lo incorpore en su estudio.

Capítulo 3

Exploración de métodos de mitigación sobre ejemplos sintéticos

En este capítulo se proponen y describen algunos métodos para la mitigación del efecto de los errores sistemáticos en la fase de entrenamiento de los algoritmos de clasificación supervisada. En particular, se busca que estos nuevos métodos sean plenamente conscientes de los errores sistemáticos que afectan a la muestra de entrenamiento. Así, los algoritmos aprenderán esta debilidad y podrán apoyarse en nuevos patrones o relaciones entre variables a la hora de efectuar la clasificación. Como fase previa al testeo de estos métodos en un caso realista, se analiza su efectividad en ejemplos sintéticos haciendo uso de la interfaz de Keras para RStudio [41].

Los métodos propuestos en este apartado se han clasificado en dos grandes grupos. Por un lado, se tienen los **métodos simplificados**, que buscan incluir el efecto de los errores sistemáticos mediante modificaciones en la función de pérdidas del método multivariante utilizado. Por otro lado, se encuentran los **métodos de réplicas**, que se basan en aumentar el conjunto de entrenamiento (técnica conocida como *data augmentation*) para abordar el problema de los sistemáticos.

3.1. Modelo de base

Antes de pasar al caso realista, se estudiaron varios ejemplos sintéticos, entre los que cabe destacar el siguiente. Se tiene un total de 40.000 puntos distribuidos aleatoriamente en torno a las superficies de dos esferas concéntricas tridimensionales de radios 8 y 10 unidades arbitrarias, respectivamente. Se asigna a cada uno de estos puntos una clase: bien señal (codificada como "1") o bien fondo (codificada como "0"), en analogía con el caso realista que se quiere estudiar. En lo que sigue, se asume que la clase de fondo es la asociada a los puntos distribuidos en torno a la superficie de la esfera de radio 8, mientras que la clase de señal es la relativa a los puntos de esfera de radio 10.

Los puntos se generan en coordenadas esféricas (donde el problema es separable), pero se transforman a coordenadas cartesianas (donde el problema es más complejo asumiendo que la simetría esférica es desconocida). La manera de generar los puntos en coordenadas cartesianas es aplicando la matriz de cambio de base desde coordenadas esféricas a un conjunto de ternas aleatorias (r,θ,ϕ), donde r representa el radio de la esfera, θ el ángulo polar y ϕ el ángulo azimutal.

En concreto, los puntos de la clase de fondo se generan transformando 20.000 ternas (r,θ,ϕ), con radio tomado de una distribución normal de media 8 y desviación típica 1, ángulo polar con distribución uniforme en el intervalo $[0, \pi]$ y el ángulo azimutal con distribución uniforme en el intervalo $[-\pi, \pi]$. Asimismo, los puntos de la clase de señal se toman con $r \sim \mathcal{N}(10, 1^2)$, $\theta \sim \mathcal{U}[0, \pi]$ y $\phi \sim \mathcal{U}[-\pi, \pi]$. Las distintas observaciones se representan en la Figura 3.1.

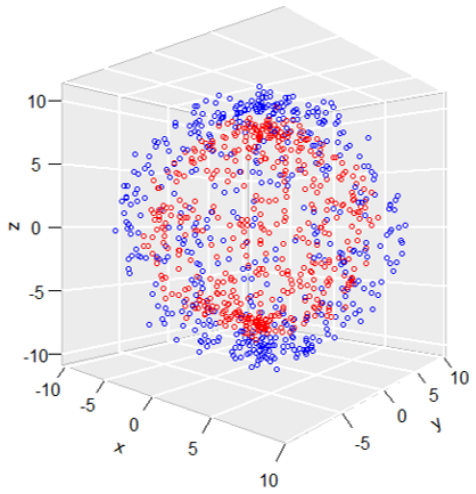

Figura 3.1: Representación tridimensional del conjunto de datos. Los puntos azules se corresponden con la clase de señal, mientras que los rojos están asociados a la clase de fondo.

El problema de separar las clases de señal y fondo en el ejemplo anterior se trata, por tanto, de un problema no lineal que los clasificadores más sencillos (p. ej. discriminante lineal, discriminante de Fisher [27]) no van a ser capaces de resolver. El propósito de los métodos expuestos en las secciones siguientes es doble. Por un lado, que el método sea capaz de resolver de una manera aceptable un problema de clasificación binaria similar al anterior. Por otro lado, que considere el efecto de los errores sistemáticos en la fase de entrenamiento para que el método sea más robusto ante nuevas observaciones.

3.2. Procedimiento general

En primer lugar, las distintas observaciones se clasifican aleatoriamente en tres grandes grupos: entrenamiento, test, y validación. La utilidad de cada uno de estos conjuntos de datos está descrita al comienzo de la Sección 2.1.

Las observaciones anteriores son las consideradas nominales, esto es, los datos ideales, que no se ven afectados por ningún tipo de error sistemático. Ahora bien, dado que nuestro objetivo es incorporar estos últimos en la fase de entrenamiento, es preciso considerar cómo se ven afectados nuestros datos ideales tras incluir el efecto de los errores sistemáticos en las variables. En consecuencia, para cada uno de los conjuntos de entrenamiento, test y validación, además del nominal, tendremos un conjunto con la sistemática *up* (con el efecto positivo del sistemático) y un conjunto con la sistemática *down* (con el efecto negativo del sistemático) que se obtienen como se indica en la Sección 2.2.1. Por ejemplo, el efecto sistemático podría consistir en que todos los sucesos tengan la variable x desplazada un 5 % de su valor.

El procedimiento general que se ha seguido para evaluar la efectividad de los métodos propuestos consiste en hacer una comparativa entre las siguientes alternativas:

CASO I Se encuentra un modelo de base aceptable (véase la Sección 2.1.3) que resuelva el problema de clasificación binaria en cuestión. Este modelo se entrena con la muestra de entrenamiento nominal, y su rendimiento se determina con la muestra de test nominal. A continuación, se evalúa dicho modelo en la muestra de validación nominal y se obtiene la curva ROC asociada a la clasificación.

CASO II Se evalúa el modelo de base obtenido en el paso anterior (entrenado con la muestra nominal, sin sistemáticos) en la muestra de validación con la sistemática *up* (respectivamente sistemática *down*), y se obtiene la curva ROC correspondiente.

CASO III Se entrena un nuevo modelo que incluya el efecto de los errores sistemáticos (según las especificaciones detalladas en las Secciones 3.3 y 3.4). A continuación, se evalúa el nuevo modelo en la muestra de validación con la sistemática *up* (respectivamente sistemática *down*) y se obtiene la curva ROC de la clasificación. Lo que se propone en este trabajo es, precisamente, encontrar un método que tenga en cuenta los sistemáticos cuyo rendimiento sea mejor que el obtenido en el CASO II.

Con el objetivo de comparar los tres casos anteriores, se representan en la misma gráfica las curvas ROC correspondientes a cada una de las tres clasificaciones efectuadas (similar a la Figura 3.2). Se tendrá una gráfica para los modelos evaluados en la muestra de validación con la sistemática *up*, y otra distinta para los evaluados en la muestra con la sistemática *down*.

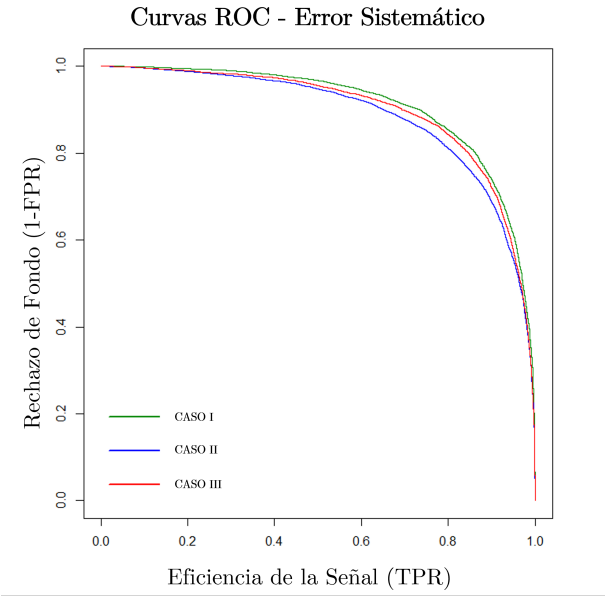

Figura 3.2: Comportamiento esperado de las curvas ROC al tratar de mitigar el efecto de los errores sistemáticos con alguno de los métodos propuestos en este capítulo. La línea verde (CASO I) representa la curva ROC de una situación ideal en la que el modelo de base se evalúa en la muestra de validación nominal, mientras que la línea azul (CASO II) se corresponde con la curva ROC al evaluar en el modelo de base la muestra de validación con la sistemática. La línea roja (CASO III) muestra la curva ROC resultante de evaluar la muestra de validación con la sistemática en el modelo entrenado teniendo en cuenta los errores sistemáticos.

El modelo de base evaluado en la muestra de validación nominal representaría la situación ideal, esto es, el modelo se entrena con datos ideales y se evalúa sobre datos también ideales. Por otro lado, evaluar el modelo de base en la muestra de validación con la sistemática *up* (respect. *down*) representa lo que ocurre en un análisis real. En efecto, en un caso realista el algoritmo se entrena con muestras MC ideales y, a continuación, se evalúa en datos reales afectados por errores sistemáticos no tenidos en cuenta en el proceso de entrenamiento. Finalmente, si al evaluar el modelo entrenado teniendo en cuenta sistemáticos en la muestra de validación con la sistemática *up* (respect. *down*) se observa una mejora con respecto al Caso II, se podría conjeturar que el método es capaz de mitigar el efecto de los errores sistemáticos a la vez que proporciona un buen rendimiento. Lo ideal sería que este nuevo método tuviese un rendimiento lo más próximo al del modelo de base evaluado en la muestra de validación nominal (Caso I).

En lo que sigue, se utilizan como método multivariante las redes neuronales artificiales. Tras un proceso de optimización de los hiperparámetros que definen las ANN, la configuración que permitía obtener un modelo de base aceptable para el ejemplo de la Sección 3.1 es la que se indica en la Tabla 3.1.

Hiperparámetros de la ANN	**Valor optimizado**
Topología	3:8:8:2
Función activación	ReLu - ReLu - Softmax
Función de Pérdidas	Error Cuadrático Medio
Optimizador	Adam
Learning rate	0.01
Epochs entrenamiento	50
Tamaño del *batch*	250

Tabla 3.1: Resumen de los hiperparámetros empleados para el entrenamiento de las redes neuronales artificiales involucradas en el análisis efectuado en este capítulo.

Finalmente, en la Figura 3.3 se muestra la curva de aprendizaje del modelo de base. En rojo se representa el valor de la función de pérdidas (MSE) en cada una de las *epochs* para la muestra de entrenamiento, mientras que en azul se hace lo mismo para la muestra de test.

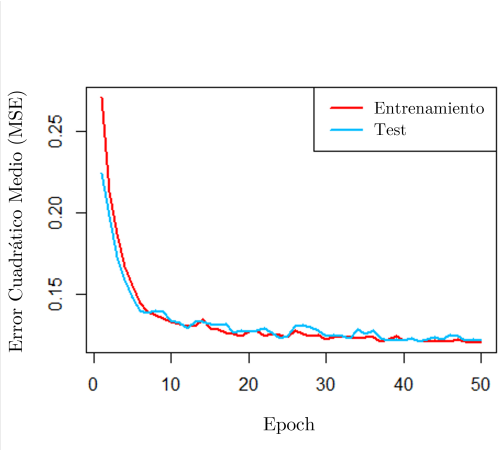

Figura 3.3: Curva de aprendizaje de la ANN entrenada con los parámetros de la Tabla 3.1.

A la vista de la figura anterior, se concluye que no se ha producido un sobreajuste a la hora de entrenar el modelo de base. La razón es que la función de pérdidas en las muestras de entrenamiento y de test tienen un comportamiento decreciente prácticamente similar. Un claro indicador de sobreajuste del algoritmo sería que el error en la muestra de test incrementase conforme disminuye el de la muestra de entrenamiento.

3.3. Métodos simplificados

La filosofía de los métodos simplificados consiste en colapsar el efecto de todos los errores sistemáticos en una única variable. Los modelos producidos a partir de estos métodos se obtienen entrenando el algoritmo multivariante después de introducir una serie de modificaciones en la función de pérdidas. En concreto, la contribución de cada dato del conjunto de entrenamiento a la función de error se corrige por un peso que se calcula en base a esta variable que condensa toda la información.

En consecuencia, se trata de una técnica con menos complejidad que el *data augmentation*, pues no es necesario replicar las variables ni modificar el conjunto de entrenamiento. Por ello, de comprobarse la efectividad en la práctica de estos métodos, constituirían una solución muy simplificada y fácil de implementar para la mitigación de los errores sistemáticos.

A continuación, se proponen dos métodos simplificados que han sido objeto estudio en este trabajo: el método ρ-simplificado y el método θ-simplificado. Los nombres se deben a las variables sobre las que se condensa el efecto de los errores sistemáticos.

3.3.1. Método ρ-simplificado

El punto de partida es un modelo de base aceptable, por ejemplo el de la Sección 3.2. Los pesos que se le asignarán a cada dato del conjunto de entrenamiento en la función de pérdidas se obtendrán evaluando el modelo de base anterior en las muestras de test nominal, con la sistemática *up* y con la sistemática *down*. Por simplicidad, se va a modelizar un método que tenga en cuenta el efecto de un solo error sistemático. A modo de ilustración, podemos suponer que es un $\pm 20\,\%$ en la variable x de los datos del ejemplo de la Sección 3.1.

En lo que sigue, se expone un desarrollo matemático que motiva el uso este tipo de metodologías para la mitigación del efecto de los errores sistemáticos.

Una vez entrenado un modelo aceptable, podremos evaluar los datos del test correspondientes a la clase de señal (codificada como "1") y a la clase de fondo (codificada como "0") de manera independiente. Esto se debe a que la etiqueta o categoría de cada dato de la muestra test es conocida. Es más, podemos organizar el *output* de esta clasificación en forma de un histograma (véase la Figura 3.4).

En este caso, la expresión del error cuadrático medio discretizado para los datos del test de la categoría k, con $k \in \{0, 1\}$, se puede aproximar por

$$\mathrm{E}_k(\boldsymbol{\omega}) = \frac{1}{N_k} \sum_{i=1}^{M} \rho_{i,k} \left(\theta_i(\boldsymbol{\omega}) - t_k\right)^2 , \tag{3.1}$$

donde N_k es el número total de datos de la muestra test pertenecientes a la clase k, M el número total de *bins* en el histograma, $\rho_{i,k}$ el número de eventos de la categoría k que caen en el *bin*

i-ésimo del histograma, $\theta_i(\boldsymbol{\omega})$ es la predicción asignada a todos los datos cuyo output cae en el *bin* i-ésimo, y t_k el *output* esperado de los datos pertenecientes a la categoría k (será $t_0 = 0$ para el fondo; $t_1 = 1$ para la señal). Con esta notación, el error cuadrático medio de la muestra de test completa (es la unión disjunta de las observaciones asociadas a la clase de fondo y las asociadas a la clase de señal) será aproximadamente:

$$\mathrm{E}(\boldsymbol{\omega}) = \frac{N_0 \mathrm{E}_0(\boldsymbol{\omega}) + N_1 \mathrm{E}_1(\boldsymbol{\omega})}{N_0 + N_1} \tag{3.2}$$

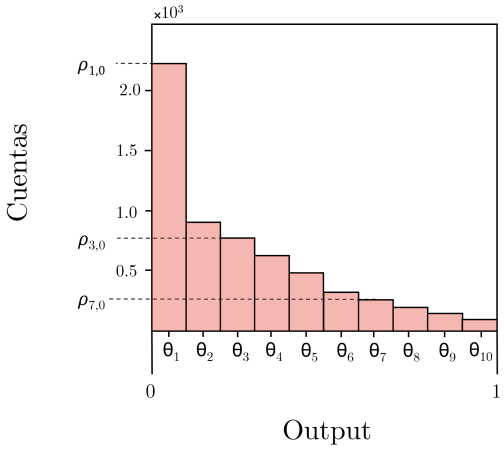

Figura 3.4: *Outputs* del modelo de base al evaluarlo en una muestra test nominal formada exclusivamente por observaciones de fondo (codificada como "0").

Supongamos que el efecto del error sistemático que queremos tener en cuenta puede modelarse mediante un parámetro que indica su magnitud, al que llamamos ε, de manera que cuando $\varepsilon = \varepsilon_0$ estaremos indicando que el error sistemático no se aplica, no existe, es la situación ideal. Por el contrario, si $\varepsilon \neq \varepsilon_0$ será un indicador de que el error sistemático "está activo". Por ejemplo, si se tiene un único error sistemático de un 5 % en la variable x, podríamos identificar la situación ideal con $\varepsilon_0 = 1$ y la situación afectada por la sistemática como $\varepsilon = 1 \pm 0.05$, en función de si se considera el efecto *up* o *down* del error.

En base a esto, podemos estudiar la dependencia funcional con este parámetro de la función de pérdidas discretizada, asumiendo que los pesos que definen la red son los mismos (el modelo de base sigue siendo el mismo). Se tiene que la cantidad $\rho_{i,k}$ en la Ec.(3.1) depende del parámetro ε porque el número de observaciones en el *bin* i-ésimo del histograma cambiará al introducir el efecto del error sistemático. Sin embargo, la cantidad $\theta_i(\boldsymbol{\omega})$ es constante para cada *bin*. En esta situación surge de forma natural al siguiente pregunta: ¿cómo afecta una variación del error sistemático a la función de pérdidas discretizada? La manera más directa de imaginárselo es a partir de una perturbación del estado de "equilibrio", lo que matemáticamente puede representarse como un desarrollo en serie de Taylor en torno a la situación ideal, i.e, cuando el parámetro que modela el error sistemático es $\varepsilon = \varepsilon_0$. Así,

$$\mathrm{E}_k(\boldsymbol{\omega}, \varepsilon) \simeq \mathrm{E}_k(\boldsymbol{\omega}, \varepsilon_0) + \left.\frac{\partial \mathrm{E}_k}{\partial \varepsilon}\right|_{\varepsilon=\varepsilon_0} (\varepsilon - \varepsilon_0), \tag{3.3}$$

donde $\mathrm{E}_k(\boldsymbol{\omega}, \varepsilon_0)$ es el error cuadrático medio al evaluar en el modelo de base las observaciones de la clase k de la muestra de test nominal. Es, por tanto, una cantidad conocida.

Tomando las derivadas parciales correspondientes y teniendo en cuenta las dependencias mencionadas con anterioridad, la Ec.(3.3) en su versión desarrollada quedaría:

$$\mathrm{E}_k(\boldsymbol{\omega}, \varepsilon) \simeq \frac{1}{N_k} \sum_{i=1}^{M} \rho_{i,k}(\boldsymbol{\omega}, \varepsilon_0) \left[1 + \frac{1}{\rho_{i,k}(\boldsymbol{\omega}, \varepsilon_0)} \frac{\partial \rho_{i,k}(\boldsymbol{\omega}, \varepsilon)}{\partial \varepsilon} \bigg|_{\varepsilon=\varepsilon_0} (\varepsilon - \varepsilon_0) \right] (\theta_i(\boldsymbol{\omega}) - t_k)^2 \,, \tag{3.4}$$

donde $\rho_{i,k}(\boldsymbol{\omega}, \varepsilon_0)$ es nuevamente el número de observaciones en el *bin* i-ésimo del histograma cuando se evalúan en el modelo de base los datos de la clase k de la muestra de test nominal.

Ahora, podemos interpretar $\frac{\partial \rho_{i,k}}{\partial \varepsilon}\big|_{\varepsilon=\varepsilon_0}$ como una variación en torno a la situación ideal al introducir el efecto del error sistemático. Por ello, podemos escribir $\frac{\partial \rho_{i,k}}{\partial \varepsilon}\big|_{\varepsilon=\varepsilon_0}(\varepsilon - \varepsilon_0) \approx \Delta\rho_{i,k}(\varepsilon)$, donde el término $\Delta\rho_{i,k}(\varepsilon)$ representa cómo cambia la variable ρ debido al efecto combinado de la sistemática *up* y *down*. Con todo ello,

$$\mathrm{E}_k(\boldsymbol{\omega}, \varepsilon) \simeq \frac{1}{N_k} \sum_{i=1}^{M} \rho_{i,k}(\boldsymbol{\omega}, \varepsilon_0) \left[1 + \frac{\Delta\rho_{i,k}(\boldsymbol{\omega}, \varepsilon)}{\rho_{i,k}(\boldsymbol{\omega}, \varepsilon_0)} \right] (\theta_i(\boldsymbol{\omega}) - t_k)^2 \,, \tag{3.5}$$

Por tanto, en forma compacta quedaría

$$\mathrm{E}_k(\boldsymbol{\omega}, \varepsilon) \simeq \frac{1}{N_k} \sum_{i=1}^{M} \beta_{i,k}(\boldsymbol{\omega}, \varepsilon, \varepsilon_0)\ \rho_i(\boldsymbol{\omega}, \varepsilon_0)\ (\theta_i(\boldsymbol{\omega}) - t_k)^2 \,, \tag{3.6}$$

En cierto modo, a falta de detallar algunos aspectos, el desarrollo anterior justifica que el efecto de los errores sistemáticos se puede modelizar mediante un repesado de la función de pérdidas. No obstante, la expresión exacta de los pesos se determinará en función de la finalidad que se tenga.

Teniendo en cuenta que el propósito del método que buscamos es mitigar el efecto de los errores sistemáticos, lo más razonable sería que aquellos datos que se vean muy afectados por la sistemática tengan pesos menores. Así, el algoritmo de optimización tenderá a ignorarlos, priorizando la correcta clasificación del resto de observaciones.

Si fuese al revés, es decir, si se asignasen pesos mayores a los datos muy afectados por los efectos sistemáticos, se estaría forzando al algoritmo a aprender a clasificar mejor las observaciones anómalas que las más "regulares"(no afectadas tan dramáticamente por los sistemáticos). Esto generaría un modelo que no respondería bien ante la mayoría de los datos.

Siguiendo esta filosofía, aunque quizá la clasificación se degrade ligeramente[1], parece natural que una manera de definir los pesos $\beta_{i,k}(\boldsymbol{\omega}, \varepsilon, \varepsilon_0)$ de la Ec.(3.6) es la siguiente:

$$\beta_{i,k}(\boldsymbol{\omega}, \varepsilon, \varepsilon_0) := \frac{1}{1 + \dfrac{\delta\rho_{i,k}(\boldsymbol{\omega}, \varepsilon)}{\rho_{i,k}(\boldsymbol{\omega}, \varepsilon_0)}}, \text{ con } k = 1, 2. \tag{3.7}$$

[1] se recuerda que el propósito de este método no es tener una excelente precisión en la clasificación, sino maximizar la significación estadística de la señal o minimizar los eventos de fondo mal clasificados como señal.

donde $\delta\rho_{i,k}(\boldsymbol{\omega}, \varepsilon)$ representa la variación en el número de observaciones en el bin i-ésimo del histograma para los datos de la clase k como consecuencia del efecto de los errores sistemáticos.

La cuestión ahora es: ¿qué expresión concreta tiene la cantidad anterior? En este trabajo se propone la siguiente:

$$\delta\rho_{i,k}(\boldsymbol{\omega}, \varepsilon) := \frac{|\rho_{i,k}^{\mathrm{UP}}(\boldsymbol{\omega}, \varepsilon) - \rho_{i,k}^{\mathrm{DOWN}}(\boldsymbol{\omega}, \varepsilon)|}{2}, \tag{3.8}$$

donde $\rho_{i,k}^{\mathrm{UP}}(\boldsymbol{\omega})$ es el número de observaciones en el *bin* i-ésimo al evaluar en el modelo de base los datos de la categoría k de la muestra test con la sistemática *up*, y $\rho_{i,k}^{\mathrm{DOWN}}(\boldsymbol{\omega})$ es el equivalente con la muestra test con la sistemática *down*. En la Figura 3.5 se ilustra la manera de obtener estos valores.

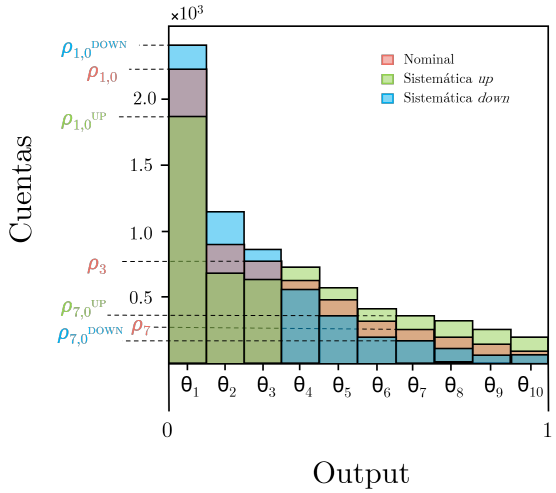

Figura 3.5: *Outputs* resultantes de evaluar en el modelo de base los datos de la clase de fondo de la muestra test nominal (color rojo), test con la sistemática *up* (color verde) y test con las sistemática *down* (color azul). Aunque los *outputs* no sean discretos, se recuerda que para este método se clasifican en *bins* (en este ejemplo, concretamente, en 10 *bins* equiespaciados). El histograma para la clase de señal se construye de manera análoga, con la salvedad de que en ese caso los *bins* más frecuentes son los más próximos a un *output* de 1.

Una vez establecida la manera de determinar la cantidad $\delta\rho_{i,k}(\boldsymbol{\omega}, \varepsilon)$, se tienen las herramientas necesarias para calcular los pesos asociados a cada uno de los datos del conjunto de entrenamiento. Para ello, se sigue el siguiente procedimiento:

1. Se evalúan los datos de la **muestra test nominal** en el modelo de base. Esta evaluación se hace por separado, según a que clase k corresponda cada dato. A continuación, se organizan los *outputs* en forma de histograma (uno distinto para cada clase) y se obtienen los valores de $\rho_{i,k}(\boldsymbol{\omega}, \varepsilon_0)$ para cada *bin* y para cada clase k. El histograma de color rojo de la Figura 3.5 ilustra lo que se esperaría obtener en esta situación.

2. Se evalúan los datos de la **muestra test con la sistemática *up*** en el modelo de base y se obtiene $\rho_{i,k}^{\mathrm{UP}}(\boldsymbol{\omega}, \varepsilon)$ para cada *bin* del histograma y cada clase. Esta situación la ejemplifica el histograma de color verde en la Figura 3.5.

3. Al igual que en el paso anterior, se evalúan los datos de la **muestra test con la sistemática *down*** en el modelo de base y se obtiene $\rho_{i,k}^{\mathrm{DOWN}}(\boldsymbol{\omega}, \varepsilon)$ para cada *bin* del histograma y cada clase. Esta situación la ejemplifica el histograma de color azul en la Figura 3.5.

4. Se calcula, para cada *bin* del histograma y cada clase k, el peso $\beta_{i,k}(\boldsymbol{\omega}, \varepsilon, \varepsilon_0)$ correspondiente dado por la Ec.(3.7) al sustituir en ella la expresión de la Ec. (3.8). Hasta aquí, cada *bin* del histograma tiene asignado un peso, en función de la clase a la que pertenezcan los datos que "caigan" en él.

5. Se evalúan los datos de la **muestra de entrenamiento** en el modelo de base y se organiza el *output* correspondiente en un histograma similar al de los pasos 1-3. Así, se consigue que cada uno de los datos del conjunto de entrenamiento caigan en algún *bin* y se le pueda asignar un peso. Supongamos que el dato j-ésimo, que pertenece a una clase k (las etiquetas son conocidas para la muestra de *training*), ha caído en el bin i-ésimo. En esta situación, al dato j-ésimo se le asigna un peso

$$\alpha_j = \beta_{i,k}(\boldsymbol{\omega}, \varepsilon, \varepsilon_0)$$

siendo $\beta_{i,k}(\boldsymbol{\omega}, \varepsilon, \varepsilon_0)$ los calculados en el paso 4. En particular, todos los datos del conjunto de entrenamiento que pertenezcan a la misma clase y hayan caído en el mismo *bin* tendrán el mismo peso asociado.

El procedimiento anterior permite obtener un peso para cada una de las observaciones del conjunto de entrenamiento, como se deseaba. En consecuencia, se está en condiciones de entrenar el nuevo modelo. Para ello, se modifica la función de pérdidas de la siguiente manera:

$$E(\hat{\boldsymbol{\omega}}) = \frac{1}{N} \sum_{j=1}^{N} \alpha_j \left(\theta_j(\hat{\boldsymbol{\omega}}) - t_j\right)^2 \tag{3.9}$$

donde N es el número total de observaciones en la muestra de entrenamiento, $\theta_j(\hat{\boldsymbol{\omega}})$ es la predicción para el dato j-ésimo, t_j, con $t_j \in \{0, 1\}$, es el valor real para el dato j-ésimo y α_j es el peso asociado a dicho dato obtenido como se ha explicado anteriormente.

Resultados del método ρ-simplificado

Se ha considerado el ejemplo de la Sección 3.1 con un único error sistemático del $\pm 20\,\%$ en la variable x de todas las observaciones. La Figura 3.6 muestra la comparativa entre curvas ROC que resulta de interés para nuestro análisis (véase la Sección 3.2). Cabe destacar que no se observa una mejora significativa con el método ρ-simplificado, como puede comprobarse de manera cuantitativa en la Tabla 3.2. De hecho, al evaluar la muestra de validación con sistemática *down* la clasificación se degrada notablemente.

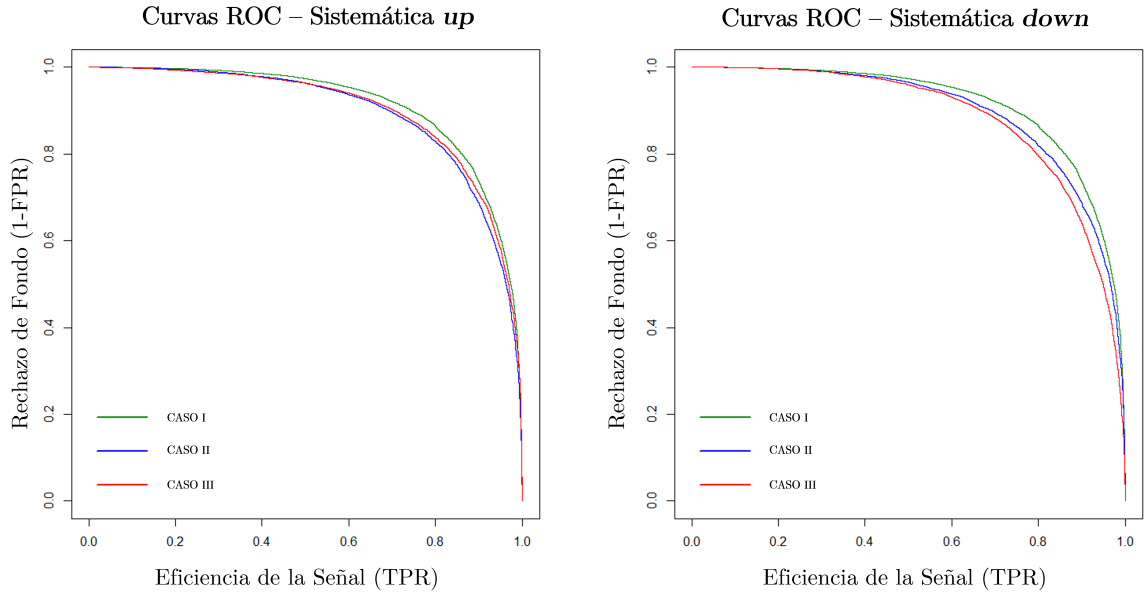

Figura 3.6: Comparativa de las curvas ROC al tratar de mitigar el efecto de un error sistemático del 20 % en la variable x usando el método ρ-simplificado en el ejemplo de la Sección 3.1. La línea verde (Caso I) representa la curva ROC de una situación ideal en la que la muestra de validación nominal se evalúa en el modelo de base, mientras que la línea azul (Caso II) se corresponde con la curva ROC al evaluar la muestra de validación con sistemática (*up* a la izquierda, *down* a la derecha) en el modelo de base. La línea roja (Caso III) muestra la curva ROC resultante de evaluar en el modelo entrenado teniendo en cuenta los errores sistemáticos la muestra de validación con la sistemática (*up* a la izquierda, *down* a la derecha).

CASUÍSTICA	***Accuracy***	**AUC**	**ES a 0.7 RF**	**ES a 0.8 RF**	**ES a 0.90 RF**	**ES a 0.95 RF**
Caso I	83.170	0.914	0.916	0.861	0.744	0.615
Caso II (*up*)	80.370	0.896	0.892	0.829	0.694	0.552
Caso III (*up*)	80.380	0.900	0.904	0.840	0.705	0.563
Variación III-II (*up*)	0.010	0.004	0.012	0.011	0.011	0.011
Caso II (*down*)	78.270	0.897	0.892	0.819	0.692	0.558
Caso III (*down*)	78.680	0.883	0.872	0.796	0.668	0.536
Variación III-II (*down*)	0.410	-0.014	-0.020	-0.024	-0.024	-0.022

Tabla 3.2: Algunos valores concretos de las métricas obtenibles a partir de las curvas ROC de la Figura 3.6. Se ha codificado como "ES" la eficiencia de la señal y "RF" el rechazo de fondo.

En base a los resultados anteriores, se observa que el método no es capaz de mitigar el efecto del error sistemático en el ejemplo considerado. Es más, se ha comprobado que tampoco se obtienen resultados competitivos al considerar otros errores sistemáticos en el mismo ejemplo, e incluso al probar en otros ejemplos sencillos distintos. En particular, se ha probado sin éxito con *spots* gaussianos y con una retícula formada por clases alternadas, tanto en el caso bidimensional como en el tridimensional.

Se concluye, por tanto, que el método ρ-simplificado no constituye una metodología eficaz que permitan mitigar el efecto de los errores sistemáticos (al menos en todos los ejemplos que

se han probado). En parte, esto puede deberse a que presenta algunas limitaciones: que la variable ρ que mide el número de observaciones en cada *bin* y define los pesos no es continua, que el entrenamiento dependerá mucho de los datos considerados porque la red no aprende la variabilidad sino que aprende a penalizar datos concretos...

3.3.2. Método θ-simplificado

En analogía con el método anterior, el método θ-simplificado consiste en condensar toda la información en una sola variable, esta vez la salida o *output* de la red. Esta cantidad la venimos denotando habitualmente por θ, razón por la cual este método recibe su nombre.

La principal diferencia con el método ρ-simplificado es que no se recurre a los *bins* de un histograma para el cálculo de los pesos que se le asigna a cada dato del conjunto de entrenamiento en la función de pérdidas, sino que ahora la variable que determina este peso es directamente el *output* de cada uno de los datos de manera individual. Además, estos pesos no se calculan empleando una muestra de test, sino que se calculan a partir de la propia muestra de entrenamiento. Esto se debe a que al hacerlo de manera individualizada, necesitaríamos una observación de test para cada dato del conjunto de entrenamiento, lo que no siempre es posible. Para evitar esto, se hace uso de la propia muestra de *training*.

Análogamente a como se hizo en el caso ρ-simplificado, una vez que tenemos entrenado un modelo de base, podemos evaluar datos de manera independiente. En concreto, podemos evaluar la propia muestra que ha sido empleada para el entrenamiento, así como sus versiones con la sistemática *up* y *down*. En esta situación, el error cuadrático medio tiene la forma

$$E(\boldsymbol{\omega}) = \frac{1}{N}\sum_{j=1}^{N}\left(\theta_j(\boldsymbol{\omega}) - t_j\right)^2, \tag{3.10}$$

donde $\theta_j(\boldsymbol{\omega})$ y t_j son, respectivamente, el *output* y el valor esperado de la observación j-ésima del conjunto de entrenamiento, con $j \in \{1 \dots N\}$.

Asumiendo nuevamente que el efecto de los errores sistemáticos se puede modelar mediante un parámetro ε, podemos estudiar cómo varía la función de pérdidas ante una perturbación de la situación ideal (cuando $\varepsilon = \varepsilon_0$) de manera similar a como se hizo para el método ρ-simplificado. En este caso, si se desarrolla la Ec.(3.3) aplicada a la función de pérdidas dada por la Ec.(3.10), resulta que

$$\mathrm{E}(\boldsymbol{\omega}, \varepsilon) \simeq \frac{1}{N}\sum_{j=1}^{N}\left[1 + \frac{2}{(\theta_j(\boldsymbol{\omega}, \varepsilon_0) - t_j)}\frac{\partial\theta_j(\boldsymbol{\omega}, \varepsilon)}{\partial\varepsilon}\bigg|_{\varepsilon=\varepsilon_0}(\varepsilon - \varepsilon_0)\right]\left(\theta_j(\boldsymbol{\omega}, \varepsilon_0) - t_j\right)^2, \tag{3.11}$$

donde $\theta_j(\varepsilon_0)$ es el *output* del dato j-ésimo del **conjunto de entrenamiento** en su versión nominal, esto es, no afectado por la sistemática.

Interpretando $\frac{\partial\theta_j}{\partial\varepsilon}\big|_{\varepsilon=\varepsilon_0}$ como una variación en torno a la situación ideal al introducir el efecto del error sistemático, podemos escribir $\frac{\partial\theta_j}{\partial\varepsilon}\big|_{\varepsilon=\varepsilon_0}(\varepsilon - \varepsilon_0) \approx \Delta\theta_j(\varepsilon)$, donde $\Delta\theta_j(\varepsilon)$ es una cantidad que representa el efecto combinado de la sistemática *up* y *down*. Con todo ello,

$$\mathrm{E}(\boldsymbol{\omega}, \varepsilon) \simeq \frac{1}{N}\sum_{j=1}^{N}\left[1 + \frac{2\Delta\theta_j(\varepsilon)}{(\theta_j(\boldsymbol{\omega}, \varepsilon_0) - t_j)}\right]\left(\theta_i(\boldsymbol{\omega}, \varepsilon_0) - t_j\right)^2. \tag{3.12}$$

Lo que de manera reducida puede representarse como:

$$\mathrm{E}(\boldsymbol{\omega}, \varepsilon) \simeq \frac{1}{N} \sum_{j=1}^{N} \eta_j(\boldsymbol{\omega}, \varepsilon, \varepsilon_0) \ \left(\theta_j(\boldsymbol{\omega}) - t_j\right)^2, \tag{3.13}$$

El desarrollo matemático anterior justifica que, aunque la variable sobre la que se proyecte el efecto de todos los errores sistemáticos cambie, la idea de introducir el efecto de la sistemática mediante un repesado en la función de pérdidas sigue siendo válida. Además, la expresión obtenida da una idea de la estructura que pueden tener los pesos en cuestión

Por los mismos motivos que en el método ρ-simplificado, se busca que los pesos $\eta_j(\boldsymbol{\omega}, \varepsilon, \varepsilon_0)$ sean menores para los datos más afectados por la incertidumbre sistemática. En concreto, se propone la siguiente forma funcional:

$$\eta_j(\boldsymbol{\omega}, \varepsilon, \varepsilon_0) := \frac{1}{1 + \dfrac{2\delta\theta_j(\boldsymbol{\omega}, \varepsilon)}{(\theta_j(\boldsymbol{\omega}, \varepsilon_0) - t_j)}} \tag{3.14}$$

donde la cantidad $\delta\theta_i(\boldsymbol{\omega}, \varepsilon)$ se calcula como:

$$\delta\theta_j(\boldsymbol{\omega}, \varepsilon) := \frac{|\theta_j^{\mathrm{UP}}(\boldsymbol{\omega}, \varepsilon) - \theta_j^{\mathrm{DOWN}}(\boldsymbol{\omega}, \varepsilon)|}{2}, \tag{3.15}$$

siendo $\theta_j^{\mathrm{UP}}(\boldsymbol{\omega})$ el *output* que resulta de evaluar el dato j-ésimo de la muestra de entrenamiento con sistemática *up* en el modelo de base, y $\theta_j^{\mathrm{DOWN}}(\boldsymbol{\omega})$ es el equivalente con la muestra con la sistemática *down*.

En consecuencia, la función de pérdidas a minimizar en el método θ-simplificado es similar a la dada en la Ec.(3.9), donde se toma $\alpha_j = \eta_j(\boldsymbol{\omega}, \varepsilon, \varepsilon_0)$.

Se han realizado pruebas con varios ejemplos sintéticos y los resultados obtenidos no han sido positivos. Al igual que ocurría con el método ρ-simplificado, incluso en algunas ocasiones la clasificación se degrada bastante. En consecuencia, no se incluyen las gráficas en la memoria, pues no aportan ningún resultado significativo.

En contraposición a lo que ocurría en el método ρ-simplificado, la variable sobre la que se condensa ahora el efecto de todos los errores sistemáticos sí que es continua. No obstante, la manera tan individualizada de calcular los pesos en este método (uno para cada dato de la muestra de *training*) hace que los modelos obtenidos a partir de él quizá estén sobreajustados para los datos utilizados en el entrenamiento.

3.4. Métodos de *Data Augmentation*

Las técnicas de *data augmentation* consisten en aumentar la muestra de entrenamiento añadiendo réplicas ligeramente distorsionadas de los datos ya disponibles. En el marco de este trabajo, estas pequeñas distorsiones representarán el efecto de los errores sistemáticos. La utilidad de estas técnicas reside en que tienen un gran efecto regularizador, esto es, producen modelos más simples que generalizan mejor a nuevas observaciones. Además, ayudan a reducir la probabilidad de sobreentramiento [42].

En este trabajo se han estudiado dos maneras implementar técnicas de *data augmentation* para la mitigación del efecto de los errores sistemáticos en la fase de entrenamiento: una más tradicional y una propuesta novedosa.

3.4.1. Método "tradicional"

El adjetivo tradicional para referirnos a este método se debe a que este procedimiento ya ha sido objeto de estudio con anterioridad. En efecto, esta técnica fue propuesta por Louis Lyons en 1992 [43] y ha sido analizada recientemente en [44].

El método en cuestión consiste en hacer K réplicas ligeramente distorsionadas de cada una de las observaciones del conjunto de entrenamiento. El número de réplicas es, a priori, un parámetro libre a optimizar en el modelo. La distorsión es aleatoria, aunque acorde a la distribución característica de los errores sistemáticos considerados. Así, por ejemplo, si se tiene una variable con un error sistemático del $\pm 20\,\%$ que sigue una distribución uniforme, la distorsión es un multiplicador tomado de la distribución $\mathcal{U}$[-0.2,+0.2]. Otra característica de este método es que condensa en cada réplica el efecto de varios errores sistemáticos de manera simultánea, por ejemplo, un $\pm 5\,\%$ en la variable x, y un $\pm 9\,\%$ en la variable y.

En este trabajo se ha considerado esta misma técnica, con una ligera modificación a la hora de escoger el número de réplicas y efectuar la distorsión de las observaciones. Se ha replicado cada observación tantas veces como errores sistemáticos se hayan considerado en el análisis. Además, cada una de estas réplicas se distorsiona acorde a un único sistemático, de manera independiente y aleatoria. Es decir, si se tiene un error del $\pm 5\,\%$ en la variable x y un $\pm 9\,\%$ en la variable y, se añade una réplica de la observación en cuestión cuya coordenada x se ha distorsionado (dejando la variable y intacta), y se añade otra réplica del dato cuya coordenada y se ha distorsionado (dejando la variable x intacta). La manera de hacer estas distorsiones es aleatoria, pero teniendo en cuenta la magnitud y la distribución propia del error. La razón de seguir este procedimiento es porque se espera que de esta manera el clasificador sea capaz de aprender de una manera más clara la variabilidad que presentan algunas variables como consecuencia del efecto de los errores sistemáticos.

No obstante, una de las desventajas que presenta esta técnica es que requiere modificar las variables involucradas en el estudio una a una, y de manera diferente para cada observación del conjunto de entrenamiento en función del efecto sistemático considerado y su distribución característica. Esta tarea, que en el ejemplo sintético que venimos manejando no resulta demasiado complicada, supone un gran esfuerzo añadido en un caso realista. En efecto, cada error sistemático tiene su propia distribución (p. ej. gaussiana, uniforme, log-normal [45]) y, además, habría que propagar su efecto de unas variables a otras. Asimismo, cuando el conjunto de entrenamiento es muy grande y se consideran varios errores sistemáticos en el análisis el problema se vuelve prohibitivo desde el punto de vista combinatorio y computacional.

Por ello, aunque se ha comprobado que este método permite mitigar de una manera bastante aceptable el efecto de los errores sistemáticos en el ejemplo de la Sección 3.1 (véase el Anexo A), no se comprueba en un caso realista.

3.4.2. Método reducido

El método reducido basado en réplicas que se propone en esta subsección constituye una alternativa para evitar modificar las variables una a una de manera aleatoria, como es habitual en la técnica "tradicional". En concreto, trata de resolver el problema de la implementación de los errores sistemáticos en el entrenamiento haciendo uso exclusivamente de las muestras de datos nominales, con la sistemática *up* y con las sistemática *down*. Esto es de especial interés desde un punto de vista práctico, pues en los análisis de física de altas energías estos conjuntos de

datos con la sistemática "a extremos" siempre están disponibles para evaluar el efecto de la incertidumbre. No se tiene constancia de que esta técnica haya sido utilizada con anterioridad en la bibliografía, por lo que se trata de una propuesta novedosa.

En la línea de lo comentado en secciones anteriores, se espera que esta propuesta sea menos sensible al efecto de los errores sistemáticos, esto es, que el *output* de un dato afectado por la sistemática sea similar al *output* de ese mismo dato que no se ve afectado por la misma. En otras palabras, que para cada observación i del conjunto de entrenamiento, ocurra que

$$|\theta_i(\boldsymbol{\omega}) - \theta_i^{\mathrm{SYST}}(\boldsymbol{\omega})| \longmapsto 0,$$

donde $\theta_i(\boldsymbol{\omega})$ es el *output* del dato i-ésimo nominal a su paso por el clasificador y $\theta_i^{\mathrm{SYST}}(\boldsymbol{\omega})$ es el *output* del dato i-ésimo afectado por la sistemática (bien *up*, bien *down*).

En lo que sigue, se describe la propuesta de método reducido basado en réplicas. En primer lugar, supongamos que se quiere mitigar el efecto de un solo error sistemático, por ejemplo un $\pm 20\,\%$ en la variable x del ejemplo que venimos manejando habitualmente. El método reducido consiste en aumentar el conjunto de entrenamiento de manera que, además de la muestra nominal, se incluyan las muestras de entrenamiento con la sistemática *up* y con la sistemática *down* asociadas al error en consideración. Es decir, que en lugar de tomar un multiplicador aleatorio de la distribución que sigue el error sistemático para distorsionar las observaciones como en el método tradicional, cada ejemplo de entrenamiento se distorsiona exactamente con los dos extremos del intervalo de incertidumbre correspondientes al sistemático implementado. Por ello, en el ejemplo que estamos considerando, la muestra con la sistemática *up* sería la muestra nominal en la que los valores de la variable x se ven aumentados un $20\,\%$, mientras que la muestra con sistemática *down* será la muestra nominal en la que los valores que toma la variable x han disminuido un $20\,\%$. Así, el número de muestras destinadas al entrenamiento del modelo se triplica cuando se considera el efecto de un solo error sistemático.

Una gran ventaja que disfruta este método y no hacen otras propuestas de la bibliografía [8], es que esta metodología se puede extender de una manera muy sencilla a situaciones en las que se quieren incluir los efectos de varios errores sistemáticos de manera combinada. Supongamos que estamos interesados en incluir K errores sistemáticos en nuestro análisis. Para cada uno de ellos, podemos calcular de manera independiente la muestra de entrenamiento afectada por la sistemática *up* (con el efecto positivo del error sistemático en cuestión) y por la sistemática *down* (con el efecto negativo). La metodología es la misma: incluir en el conjunto de entrenamiento, además de la muestra nominal, las muestras distorsionadas de manera independiente por los efectos extremos de los distintos errores sistemáticos considerados. En consecuencia, si la muestra de entrenamiento originalmente tiene N observaciones, al seguir el método reducido con K errores sistemáticos pasará a tener $N\,(1 + 2K)$.

En definitiva, el método reducido permite entrenar un gran modelo que tenga en cuenta el efecto de varios errores sistemáticos de manera conjunta, y cuyo fundamento es "enseñar" al algoritmo la máxima variabilidad que puede haber en las variables usadas como *inputs* en la fase de entrenamiento. Con esto, se espera que el clasificador sea capaz de encontrar nuevos patrones y se apoye en variables más robustas para efectuar la separación. Cabe mencionar que esta técnica es bastante factible desde un punto de vista práctico, pues tan solo hace uso de las muestras nominal, con la sistemática *up* y con la *down* (disponibles en la mayoría de los análisis realistas), sin necesidad de modificar aleatoriamente las variables una a una en cada observación y evitando el colapso combinatorio.

Resultados del método reducido basado en réplicas

En todo lo que sigue, nos referiremos al ejemplo de la Sección 3.1. En primer lugar, estudiaremos cómo se comporta este método simplificado con un solo error sistemático. A diferencia de otras ocasiones, vamos a suponer que dicho error consiste en una distorsión del centro de las esferas que definen el conjunto de observaciones. En particular, los centros de éstas podrán desplazarse por una de las diagonales del cubo donde están inscritas, oscilando entre los valores $\pm(1, 1, 1)$.

Se ha entrenado un nuevo modelo usando el método reducido basado en réplicas para incluir el efecto del sistemático anterior. En concreto, la muestra de entrenamiento es una concatenación de la muestra nominal, la muestra con sistemática *up* (el centro desplazado hasta (1,1,1)) y la sistemática *down* (el centro desplazado hasta (-1,-1,-1)). En la Figura 3.7 se compara el rendimiento de este modelo con el modelo de base a través de las curvas ROC.

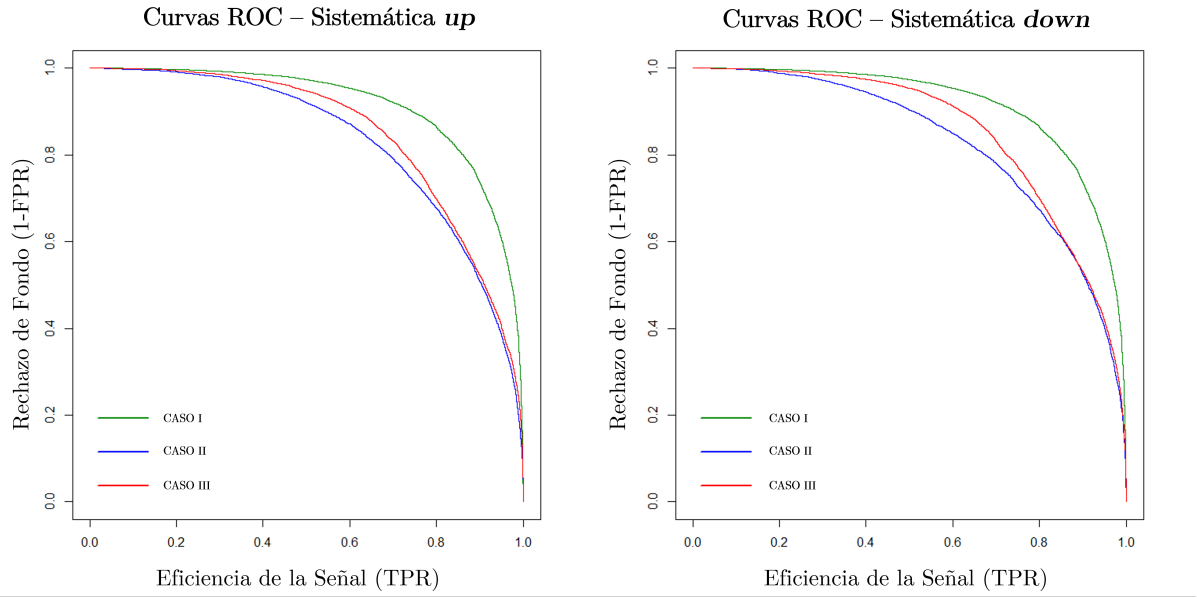

Figura 3.7: Comparativa de las curvas ROC al tratar de mitigar el efecto de un error sistemático consistente en un cambio de centro de las esferas mediante el método reducido de *data augmentation*. El significado de la leyenda es similar al de otras ocasiones (ver Figura 3.6).

CASUÍSTICA	*Accuracy*	AUC	ES a 0.7 RF	ES a 0.8 RF	ES a 0.90 RF	ES a 0.95 RF
Caso I	83.170	0.914	0.916	0.861	0.744	0.615
Caso II (*up*)	74.425	0.831	0.781	0.691	0.545	0.422
Caso III (*up*)	76.615	0.850	0.798	0.728	0.614	0.489
Variación III-II (*up*)	2.190	0.019	0.017	0.038	0.069	0.067
Caso II (*down*)	74.000	0.826	0.778	0.675	0.509	0.382
Caso III (*down*)	76.305	0.852	0.799	0.724	0.620	0.515
Variación III-II (*down*)	2.305	0.026	0.022	0.049	0.111	0.134

Tabla 3.3: Algunos valores concretos de las métricas obtenidas a partir de las curvas ROC de la Figura 3.7. Se ha codificado como "ES" la eficiencia de la señal y "RF" el rechazo de fondo.

Se observa que el modelo entrenado teniendo en cuenta el efecto del error sistemático (línea roja,

Caso III) logra una mejora notable de rendimiento con respecto al modelo de base evaluado en la muestra con sistemática (línea azul, Caso II) . La Tabla 3.3 resume de manera cuantitativa las mejoras obtenidas para ciertos puntos de interés en la ROC.

Se observa que, para un nivel de rechazo de fondo del 95 %, la eficiencia de la señal aumenta drásticamente en 13 puntos porcentuales para la sistemática *down*, y en 6 puntos para la sistemática *up*. En promedio, esto supone una mejora de más de 9 puntos con respecto al Caso II, lo que pone de manifiesto la efectividad del método reducido. Otra forma de leer esta información es en términos de reducción de errores sistemáticos. Por ejemplo, a un nivel de rechazo de fondo del 95 %, el efecto del error sistemático *down* es disminuir la eficiencia de la señal de un 61.5 % a un 38.2 %. No obstante, al utilizar el método reducido, la eficiencia de la señal aumenta hasta el 51.5 %. Esto supone una ganancia de aproximadamente 13 puntos porcentuales en la eficiencia de la señal, o equivalentemente, una reducción del efecto del error sistemático en un 43 %. Por otro lado, se puede comprobar que esta técnica siempre proporciona una ganancia, independientemente de la región de la curva ROC analizada (al menos para este ejemplo).

A continuación, se ha estudiado la efectividad del método en un ejemplo con dos errores sistemáticos: uno del ±20 % en la variable x, y otro del ±15 % en la variable y. Una vez entrenado el nuevo modelo que tiene en cuenta el efecto de estos dos errores, analizamos su rendimiento. Llegados a este punto, se pueden hacer los dos análisis siguientes:

1. Estudiar la efectividad del método al evaluarlo en una muestra afectada simultáneamente por los dos errores sistemáticos usados en el entrenamiento.
2. Estudiar el comportamiento cuando se considera el efecto de un único error.

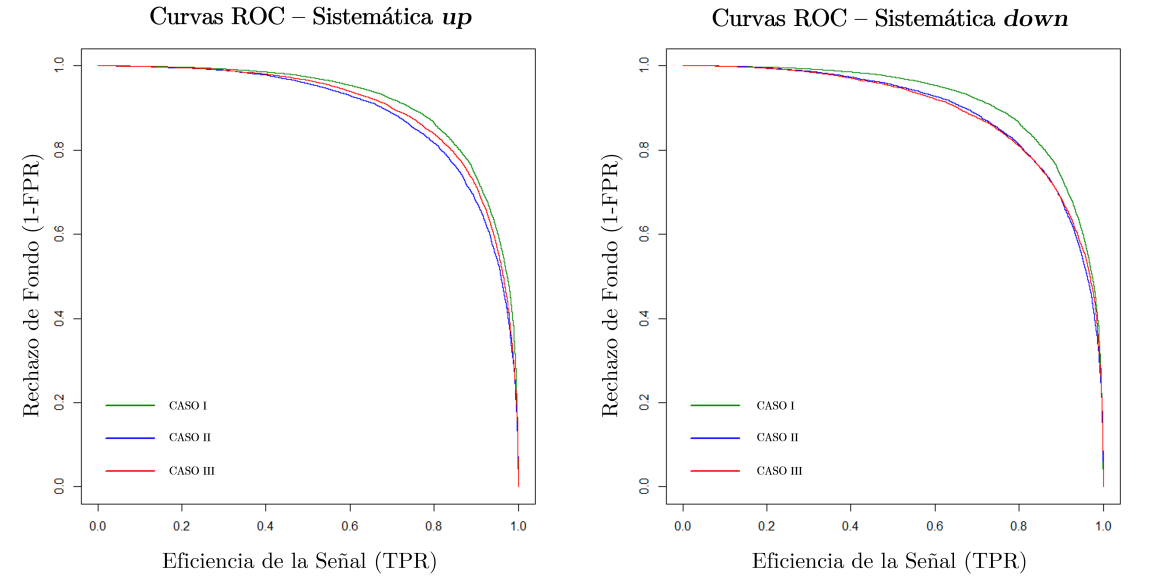

Figura 3.8: Comparativa de las curvas ROC al tratar de mitigar simultáneamente el efecto de un error sistemático del ±20 % en la variable x, y un ±15 % en la variable y mediante el método reducido de *data augmentation*.

En primer lugar, se exponen los resultados obtenidos para la primera de las situaciones anteriores. Para ello, es útil la representación de la Figura 3.8. En ella se observa que, para el caso de

una muestra afectada simultáneamente por la sistemática *up* de los dos errores sistemáticos, el modelo entrenado empleando el método de réplicas reducido proporciona una mejora notable. En efecto, la curva ROC alcanza un punto intermedio entre el Caso II (situación realista) y el Caso I (situación ideal) . Por el contrario, no hay mejora si se considera el efecto de la sistemática *down* (aunque tampoco se observa una degradación importante como ocurría en el método ρ-simplificado). Esto se puede ver cuantitativamente en la Tabla 3.4

CASUÍSTICA	***Accuracy***	**AUC**	**ES a 0.7 RF**	**ES a 0.8 RF**	**ES a 0.90 RF**	**ES a 0.95 RF**
Caso I	83.170	0.914	0.916	0.861	0.744	0.615
Caso II (*up*)	77.765	0.890	0.888	0.817	0.676	0.532
Caso III (*up*)	76.735	0.900	0.906	0.841	0.699	0.565
Variación III-II (*up*)	-1.030	0.010	0.018	0.023	0.022	0.034
Caso II (*down*)	72.765	0.890	0.891	0.811	0.670	0.518
Caso III (*down*)	74.725	0.890	0.892	0.809	0.653	0.506
Variación III-II (*down*)	1.960	0.000	0.001	-0.002	-0.016	-0.012

Tabla 3.4: Algunos valores concretos de las métricas obtenidas a partir de las curvas ROC de la Figura 3.8. Se ha codificado como "ES" la eficiencia de la señal y "RF" el rechazo de fondo.

A continuación, se muestran los resultados para la segunda situación. En ella, el modelo ha sido entrenado teniendo en cuenta el efecto de los dos errores sistemáticos, pero solamente lo evaluamos con uno de ellos. En particular, los resultados que se presentan (véase la Figura 3.9) se corresponden con una muestra afectada exclusivamente por la sistemática en la variable x.

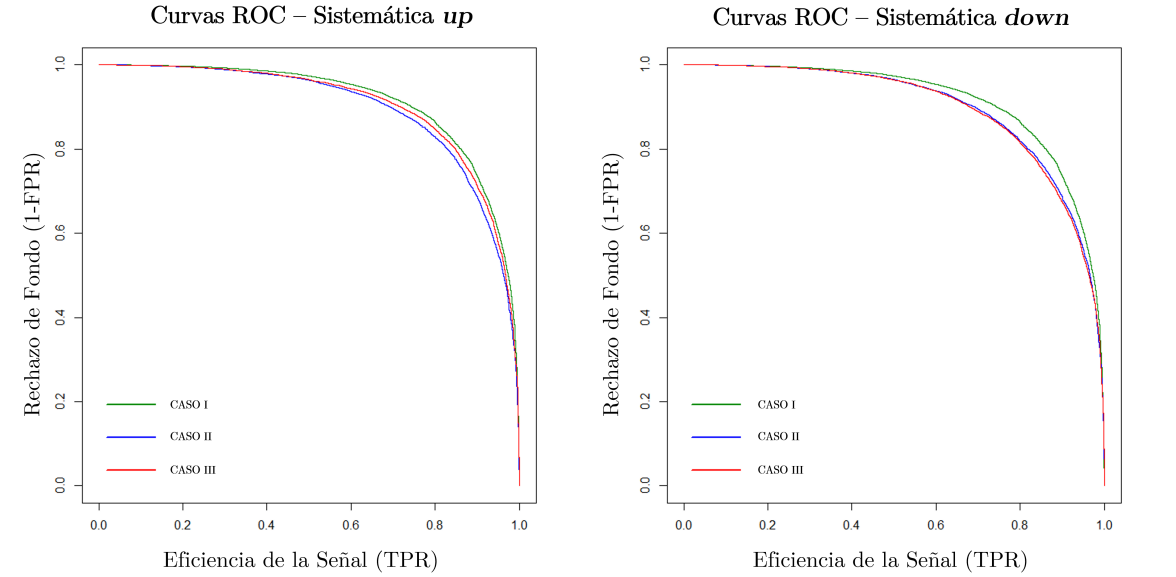

Figura 3.9: Comparativa de las curvas ROC al tratar de mitigar el efecto de un error sistemático del $\pm 20\,\%$ en la variable x al utilizar el método reducido de *data augmentation* entrenado con el efecto de un error sistemático del $\pm 20\,\%$ en la variable x, y un $\pm 15\,\%$ en la variable y. Se representa la situación en la que el modelo se entrena teniendo en cuenta el efecto de dos errores sistemáticos y su rendimiento se evalúa considerando solamente uno de ellos.

Se observa que, al considerar un solo error sistemático, el Caso II se degrada menos que al

considerar el efecto simultáneo de dos de ellos. Incluso en esta situación en la que el margen de mejora es más restringido (la diferencia entre el Caso I y el Caso II es menor), el método reducido proporciona un modelo que consigue una mejoría notable. Nuevamente, se observa que la curva ROC de la clasificación efectuada con este modelo alcanza una situación intermedia. La Tabla 3.5 muestra cuantitativamente esta mejoría para algunos puntos de interés.

Al igual que ocurría en el caso anterior, estas mejoras se obtienen para la sistemática *up*, pero no así para la sistemática *down*. En este último caso, no hay mejoría, pero tampoco hay un empeoramiento importante como el que se tenía en los métodos simplificados basados en pesos.

CASUÍSTICA	*Accuracy*	AUC	ES a 0.7 RF	ES a 0.8 RF	ES a 0.90 RF	ES a 0.95 RF
Caso I	83.170	0.914	0.916	0.861	0.744	0.615
Caso II (*up*)	80.370	0.896	0.892	0.829	0.694	0.552
Caso III (*up*)	80.105	0.904	0.905	0.848	0.718	0.575
Variación III-II (*up*)	-0.265	0.008	0.013	0.018	0.024	0.023
Caso II (*down*)	78.270	0.897	0.892	0.819	0.692	0.558
Caso III (*down*)	79.455	0.894	0.886	0.814	0.680	0.558
Variación III-II (*down*)	1.185	-0.003	-0.006	-0.005	-0.012	0.000

Tabla 3.5: Algunos valores concretos de las métricas obtenidas a partir de las curvas ROC de la Figura 3.9. Se ha codificado como "ES" la eficiencia de la señal y "RF" el rechazo de fondo.

En vista de lo comentado a lo largo de esta sección, se concluye que el método de réplicas reducido proporciona una mejora notable a la hora de implementar el efecto de los errores sistemáticos en la fase de entrenamiento. Se ha comprobado que esta metodología permite conseguir una ganancia destacable, no solo al considerar el efecto de un único error sistemático, si no también al hacer un análisis combinando varios de ellos.

3.5. Conclusiones

En este capítulo se proponen algunos métodos para la mitigación del efecto de los errores sistemáticos durante la fase de entrenamiento de los algoritmos multivariantes. En primer lugar, se presentan los métodos simplificados. Inicialmente, estos métodos parecían prometedores por su simplicidad y su capacidad para condensar toda la información en una sola variable. No obstante, se ha comprobado en ejemplos sintéticos que no proporcionan una mejora significativa con respecto al procedimiento tradicional. Es más, se ha visto que en algunas situaciones hay un empeoramiento. En esta línea, sí que es aceptable que el método no proporcione mejoras siempre (puede haber situaciones en las que no haya diferencia entre el Caso II y el Caso III en la Sección 3.2), pero no es admisible que empeore el rendimiento de manera significativa.

En segundo lugar, se han estudiado métodos de *data augmentation*. Por un lado, se ha comprobado que los métodos "tradicionales" responden bien en ejemplos sencillos. Sin embargo, la complejidad añadida que supone replicar las variables en un caso realista nos ha hecho abandonar esta técnica. Con el objetivo de evitar precisamente esto último, se propone un método de réplicas reducido que tan solo hace uso de la muestra *training* nominal, y sus variantes afectadas por la sistemática *up* y la *down*. Este método, aunque sencillo, se ha comprobado permite mitigar de una manera notable el efecto de los errores sistemáticos, al menos en casos sintéticos.

Capítulo 4

Mitigación del efecto de errores sistemáticos en un caso realista

En este capítulo, se aplica el método reducido de *data augmentation* expuesto en la Sección 3.4.2 a un caso real, en particular a un estudio sobre la producción de materia oscura en asociación con quarks top, presentado recientemente en la tesis doctoral [46]. En lo que sigue, se utiliza el análisis efectuado en dicha tesis, al que se añaden los efectos de la incertidumbre sistemática en el entrenamiento de los métodos multivariantes.

4.1. Caso de estudio: el conjunto de datos

En primer lugar, cabe mencionar que los datos empleados para entrenar los métodos multivariantes proceden de simulaciones MC, en particular de la campaña de simulaciones del verano de 2016. El tipo de señal considerado en este análisis es el proceso $t/\bar{t}$+DM. El mediador puede ser escalar o pseudoescalar, siendo su masa desconocida. Como el propósito de este capítulo no es efectuar un análisis exhaustivo de la búsqueda de materia oscura, sino comprobar la efectividad del método de reducido de *data augmentation* en un caso realista, se ha asumido que el mediador es escalar y tiene una masa de 500 GeV. Se ha tomado esta masa tan elevada porque es la situación en la que la discriminación por MVA es más importante, al ser el límite de la sensibilidad. En consecuencia, afinar el efecto de los errores sistemáticos en este límite es crucial.

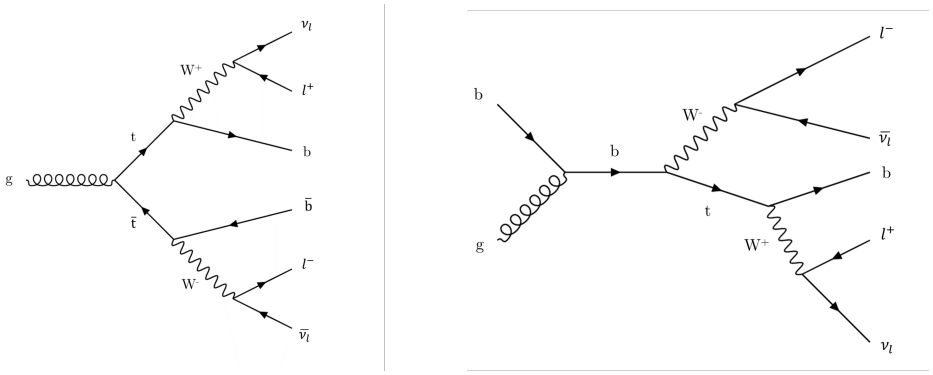

Figura 4.1: Diagramas de Feynman típicos para los procesos *ttbar* (izquierda) del SM y *single top* (derecha) del SM.

El proceso $t/\bar{t}$+DM se caracteriza porque en él se produce un par top-antitop que da lugar a un mediador de materia oscura (véase la Figura 1.2). Uno de los quarks top procede de un gluon energético, mientras que el otro procede del decaimiento de un quark b. No obstante, existen numerosos procesos físicos de fondo del SM que enmascaran esta señal. Entre ellos, los que más se asemejan son la producción aislada de quarks top, conocido como *single top* o $t/\bar{t}$(SM), y la producción de pares quark-antiquark top, conocido como *ttbar* o $t\bar{t}$(SM). En la Figura 4.1 se muestran los diagramas de Feynman típicos para estos procesos. Por simplicidad, nos centraremos en el fondo *single top* en lo que resta de capítulo. Por ello, de ahora en adelante la señal la componen sucesos $t/\bar{t}$+DM, mientras que el fondo estará formado por sucesos $t/\bar{t}$(SM).

Variable	Significado
`METcorrected_pt`	Energía transversa faltante corregida para mitigar la dependencia con el ángulo azimutal ϕ
`mt2ll`	"Pseudomasa transversa" del par de leptones
`dphillmet`	Diferencia en el plano transverso entre el momento faltante y el sistema leptón-leptón
`mblt`	Observable definido para seleccionar eventos compatibles con dos decaimientos de quarks top al estado semileptónico
`mt2bl`	"Pseudomasa transversa" del par b-jet/leptón
`massT`	Suma escalar de la componente transversa del momento faltante, los dos leptones y los dos b-jets obtenidos en la reconstrucción del quark top
`reco_weight`	Peso de reconstrucción en el proceso de reconstrucción de pares $t\bar{t}$
`cosphill`	Ángulo azimutal del sistema leptón-leptón
`costhetall`	Ángulo polar del sistema leptón-leptón
`dark_pt`	Estimación del momento transverso del mediador de materia oscura en la interacción
`overlapping_factor`	Factor de superposición de las elipses en el proceso de reconstrucción del quark top
`r2l`	Ratio entre la energía transversa faltante y el momento transverso de los dos leptones observados
`r2l4j`	Igual que `r2l`, pero considerando también en el denominador el momento transverso de los 4 primeros jets (si existen)
`nbJet`	Número de b-jets

Tabla 4.1: Variables empleadas en el análisis, junto su significado [46]. La "pseudomasa transversa" es una forma de estimar la masa del objeto en base a la cinemática del plano transverso.

La mayor diferencia entre estos procesos es su sección eficaz. En efecto, el proceso $t/\bar{t}$+DM tienen una sección eficaz de $3.05{\cdot}10^{-4}$ pb, mientras que el proceso $t/\bar{t}$(SM) tiene una sección eficaz de 38.7 pb [46]. En otras palabras, la señal queda fuertemente enmascarada por el fondo al ser la sección eficaz de los eventos de señal cinco órdenes de magnitud menor que la de los eventos de fondo. Para tratar de reducir esta diferencia, se preselecciona una región del espacio de fases mediante cortes en algunas variables con el objetivo de disminuir notablemente la contaminación debida a los procesos físicos de fondo. En particular, cabe destacar que se seleccionan sucesos que tengan al menos un b-jet, que el par leptón-leptón tenga una "pseudomasa transversa" superior a 80 GeV, y que tengan exactamente dos leptones con cargas eléctricas opuestas. Más aún, se define una zona de señal enriquecida en procesos $t/\bar{t}$+DM mediante la aplicación de un corte adicional que consiste en considerar aquellos sucesos que tengan exactamente un jet, o dos jets y exactamente un b-jet [46].

Sin embargo, aunque los cortes anteriores son capaces de reducir la sección eficaz del fondo, la diferencia entre las magnitudes de las secciones eficaces continúa siendo importante, lo que dificulta la discriminación señal-fondo. Es a partir de este momento cuando se emplean los métodos multivariantes para llevar a cabo esta tarea de clasificación. En particular, en el proceso de discriminación entre señal y fondo por MVA llevado a cabo en [46] se utilizan distintas variables cinemáticas que se resumen en la Tabla 4.1. Para una descripción más detallada, puede consultarse [46]. Se trata, por tanto, de un problema de clasificación con datos 14-dimensionales.

4.2. Entrenamiento

Los métodos multivariantes empleados en el análisis de [46] fueron las ANNs y los BDTs. Sin embargo, en el análisis efectuado en este capítulo, nos centraremos fundamentalmente en las ANNs, en línea con la metodología seguida en el Capítulo 3.

Los entrenamientos de las ANNs se realizan de manera balanceada, precisamente, en la zona enriquecida en procesos de señal $t/\bar{t}$+DM descrita en la sección anterior. Un aspecto importante es que, durante el entrenamiento, las ANNs empleadas en [46] se exponen además de a la señal $t/\bar{t}$+DM y al fondo $t/\bar{t}$(SM), a la señal $t\bar{t}$+DM y al fondo $t\bar{t}$(SM). No obstante, esto no afecta a la hora de implementar el método reducido, pues basta con aumentar el conjunto de entrenamiento únicamente con las muestras con sistemática relativas a los procesos *single top*, y considerar exclusivamente el rendimiento del clasificador relativo a la señal $t/\bar{t}$+DM y al fondo $t/\bar{t}$(SM).

Por otro lado, en [46] se llevó a cabo un proceso de optimización de los hiperparámetros de la ANN para los datos disponibles (se resumen en Tabla 4.2). Es importante destacar que son estos hiperparámetros los que se han empleado para entrenar las ANNs tanto del modelo de base, como del modelo reducido basado en réplicas que integra el efecto de los errores sistemáticos. En este contexto, se hace notar que los resultados obtenidos en este capítulo podrían mejorarse volviendo a optimizar los parámetros de las ANNs tras la inclusión de los errores sistemáticos en el conjunto de entrenamiento, aunque esta tarea está fuera del alcance de este trabajo.

Parámetro de la ANN	**Valor optimizado**
Topología capas ocultas	20,15,15,10
Funciones de activación capas ocultas	ReLu(x4)
Funciones de activación capas salida	Softmax
Función de Pérdidas	Error Cuadrático Medio (MSE)
Optimizador	Adam
Learning rate	0.001
Epochs entrenamiento	250
Tamaño del *batch*	250

Tabla 4.2: Resumen de la optimización de los parámetros de las ANNs empleados en el análisis.

4.3. Resultados

En esta sección, se presentan los resultados obtenidos al tratar de mitigar los errores sistemáticos que afectan a la MET y a la JES (véase la Sección 2.2.2) mediante el método reducido de *data augmentation* propuesto en la Sección 3.4.2.

En este análisis, se han considerado exclusivamente los errores sistemáticos que afectan a los procesos físicos de fondo. En contraposición, los sucesos de señal se han considerado en su versión nominal, esto es, como no afectados por la sistemática. La razón es que no se tuvieron disponibles las muestras de señal afectadas por la sistemática. No obstante, cabe destacar que como la sección eficaz de los procesos de fondo es varios órdenes de magnitud mayor que la de los procesos de señal, lo interesante es mitigar los errores sistemáticos que afectan sobre todo al fondo. En consecuencia, cabe esperar que el efecto de la inclusión de los errores sistemáticos en las muestras de señal tenga un efecto positivo, aunque menor, pues el problema en este tipo de búsquedas se focaliza en las colas de la distribución del fondo (y no tanto en las colas de la distribución de señal). Teniendo en cuenta esto, es necesario hacer la siguiente redefinición:

- La muestra de entrenamiento nominal está formada por muestras MC ideales (sin errores sistemáticos), tanto de la señal como del fondo. Lo mismo ocurre con las muestras de test y validación nominales.
- La muestra de entrenamiento con los efectos de la sistemática *up* (respect. *down*) está formada por muestras MC de los sucesos de fondo afectadas por la sistemática *up* (respect. *down*). No se incluyen eventos de señal.
- La muestra de validación con los efectos de la sistemática *up* (respect. *down*) incluye muestras MC ideales de señal (nominales), además de muestras MC de los sucesos de fondo afectadas por la sistemática up (respect. *down*).

En lo que sigue, se utiliza una metodología similar a la del Capítulo 3, es decir, se comparan los tres casos descritos en la Sección 3.2.

4.3.1. Modelo de base

El modelo de base se entrena con los parámetros indicados en la Tabla 4.2. En la Figura 4.2 se muestra una comprobación de que no ha habido sobreentrenamiento. Por otro lado, en la Tabla 4.3, se incluyen las 10 variables más influyentes a la hora de efectuar la clasificación junto con un índice que refleja la importancia o poder de discriminación de cada una de ellas.

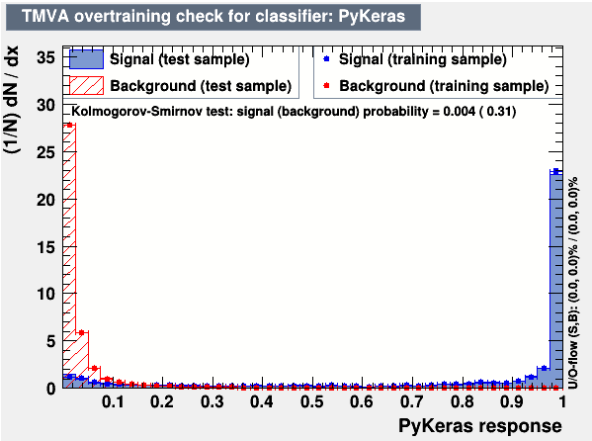

Figura 4.2: Comprobación de sobreentrenamiento del modelo de base.

Variable	Separación
mt2ll	0.6104
METcorrected_pt	0.5602
massT	0.3696
mt2bl	0.2368
r2l4j	0.2303
mblt	0.1379
r2l	0.0312
dark_pt	0.0272
reco_weight	0.0269
overlapping_factor	0.0251

Tabla 4.3: Poder de discriminación de las 10 variables más influyentes en el modeo de base.

4.3.2. Mitigación de los errores sistemáticos que afectan a la MET

En esta subsección, se tratan de mitigar los errores sistemáticos que afectan a la MET. Se ha entrenado un nuevo modelo siguiendo el método reducido de réplicas, donde se ha incluido en la fase de entrenamiento, además de la muestra de entrenamiento nominal, las muestras de entrenamiento con la sistemática *up* y *down* (véase la Sección 2.2.2). En la Figura 4.3 se comprueba que no ha habido sobreentrenamiento, mientras que en la Tabla 4.4 se muestra el poder de discriminación de las 10 variables más influyentes en la clasificación.

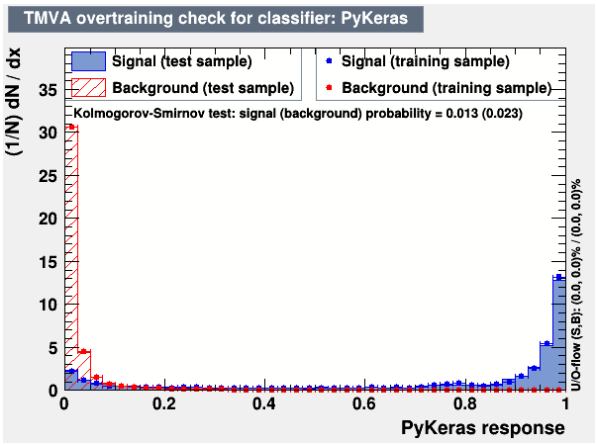

Figura 4.3: Comprobación de sobreentrenamiento del modelo reducido entrenado para mitigar los sistemáticos que afectan a la MET.

Variable	Separación
`mt2ll`	0.5913
`METcorrected_pt`	0.5004
`massT`	0.3205
`mt2bl`	0.2391
`mblt`	0.1497
`r2l4j`	0.1464
`dark_pt`	0.0255
`reco_weight`	0.0245
`overlapping_factor`	0.0234
`costhetall`	0.0230

Tabla 4.4: Poder de discriminación de las 10 variables más influyentes en el modelo reducido para mitigar los sistemáticos sobre la MET.

Comparando las Tablas 4.3-4.4, se observa que al introducir en el entrenamiento los sistemáticos que afectan a la MET, la variable `METcorrected_pt` pierde casi un 11 % de su poder de discriminación. Esto es lo esperado, pues si se está incluyendo variabilidad en la MET, lo natural es que la ANN confíe menos en esta variable.

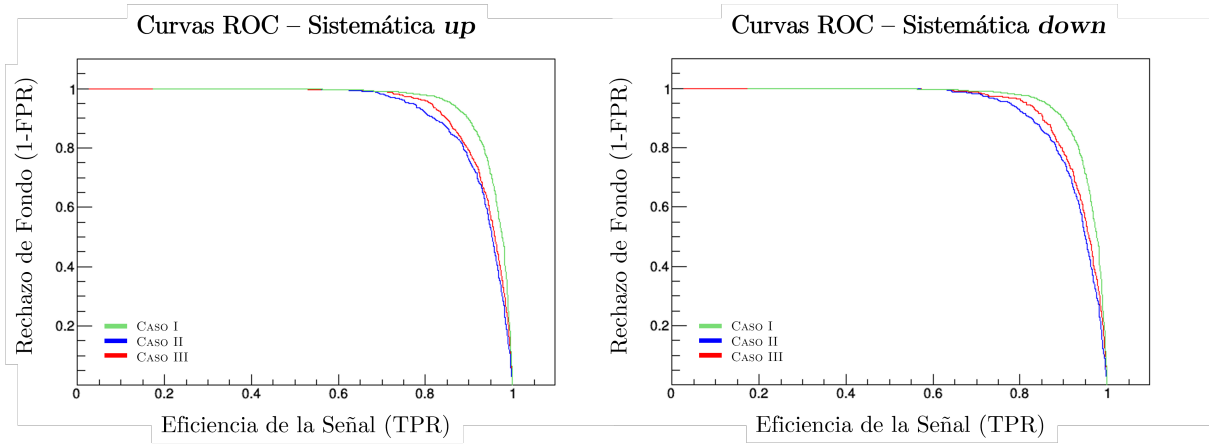

Figura 4.4: Comparativa de las curvas ROC al tratar de mitigar el efecto de los errores sistemáticos que afectan a la MET utilizando el método reducido de *data augmentation*. El significado de los distintos casos que se muestran en la leyenda se describe al comienzo de la Sección 3.2, siendo similar además a los de la Figura 3.6.

La Figura 4.4 muestra la comparativa de las curvas ROC entre el modelo de base y el modelo reducido de réplicas que tiene en cuenta los errores sistemáticos que afectan a la MET. Se observa que al evaluar en el modelo de base las muestras con incertidumbre sistemática (curva azul) el rendimiento se degrada notablemente con respecto a la situación ideal (curva verde). Sin embargo, al considerar el modelo entrenado con sistemáticos (curva roja), el rendimiento mejora. En particular, esta mejora es más pronunciada en la región superior de la ROC.

Esto último es especialmente interesante, ya que esta es la zona de trabajo habitual en experimentos de búsqueda de materia oscura. En efecto, en la realidad se tienen muchos más eventos de fondo que de señal, de manera que si el fondo no se identifica correctamente, la posibilidad de discriminar la señal por encima del fondo disminuye drásticamente. Por ello, interesan las regiones que permiten conseguir un alto nivel de rechazo de fondo, como por ejemplo el punto en el que el rechazo de fondo es del 99 %. En particular, mitigar el efecto de los errores sistemáticos en estas zonas podría marcar la diferencia entre discriminar señal o no. En la Tabla 4.5 se muestran algunos valores cuantitativos en torno a esta zona de interés.

CASUÍSTICA	ES a 0.95 RF	ES a 0.99 RF	RF a 0.80 ES	RF a 0.95 ES	RF a 0.99 ES
Caso I	0.8591	**0.7043**	**0.9780**	0.7134	0.2835
Caso II (*up*)	0.7688	**0.6662**	**0.9196**	0.5263	0.1520
Caso III (*up*)	0.8120	**0.7039**	**0.9591**	0.5731	0.2047
Efecto Original	0.0904	**0.0381**	**0.0584**	0.1871	0.1315
Efecto Mitigado	0.0472	**0.0004**	**0.0189**	0.1403	0.0789
Ganancia (*up*)	0.0432	**0.0377**	**0.0395**	0.0468	0.0526
Reducción (%)	47.8	**98.9**	**67.6**	25.0	40.0
Caso II (*down*)	0.7759	**0.6434**	**0.9243**	0.4984	0.1332
Caso III (*down*)	0.8141	**0.6821**	**0.9638**	0.5641	0.2072
Efecto Original	0.0832	**0.0610**	**0.0537**	0.2151	0.1503
Efecto Mitigado	0.0451	**0.0222**	**0.0142**	0.1493	0.0763
Ganancia (*down*)	0.0381	**0.0387**	**0.0395**	0.0658	0.0740
Reducción (%)	45.9	**63.5**	**73.5**	30.6	49.2

Tabla 4.5: Resultados cuantitativos de la mitigación de los errores sistemáticos que afectan a la MET. Con "efecto original" nos referimos a la degradación que se produce al pasar del Caso I al Caso II, mientras que el "efecto mitigado" es la degradación del Caso I al Caso III. La ganancia es la diferencia entre el "efecto original" y el "efecto mitigado".

Se observa en la Tabla 4.5 que el método reducido es capaz de mitigar el efecto de los errores sistemáticos en todos los puntos de la ROC analizados. Un punto de la ROC donde la mejora es notable es cuando la eficiencia de la señal es del 80 %. En efecto, en la situación ideal en este punto se clasificarían mal aproximadamente el 2 % de los sucesos de fondo. Sin embargo, el efecto de los sistemáticos hace que en realidad sean casi cuatro veces más el número de eventos de fondo mal clasificados (un 8 %). Ahora bien, con el método reducido que se propone en este trabajo, el porcentaje de fondo mal clasificado disminuye hasta el 4 %, lo que permite reducir el efecto de los errores sistemáticos en aproximadamente un 70 %. Por otro lado, cuando se rechaza correctamente el 99 % del fondo, el método es capaz de mitigar casi por completo (lo reduce en un 98.9 %) el efecto *up* del error sistemático, consiguiendo una mejora de casi un 3.8 % en la eficiencia de la señal.

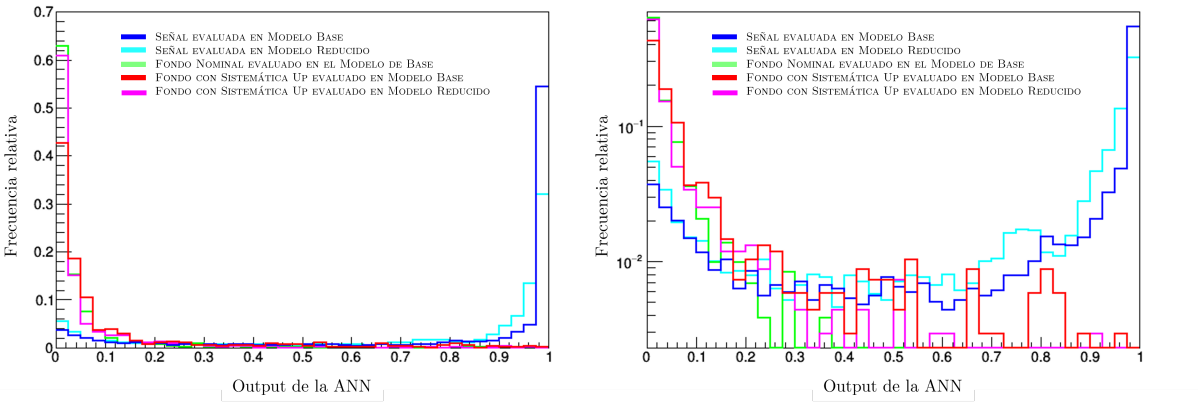

Figura 4.5: Histogramas normalizados con los *outputs* de las ANNs al tratar de mitigar el efecto sistemático que afecta a la MET. En la figura de la izquierda se representan en escala lineal, mientras que en la figura de la derecha se representan en escala logarítmica.

En la Figura 4.5 se representan los histogramas con los *outputs* resultantes de evaluar las muestras nominal y con sistemática en los distintos modelos. Por simplicidad, se incluyen los histogramas asociados a la sistemática *up*, pero los de la sistemática *down* ilustraban resultados similares. En la subfigura de la izquierda, se observa cómo al evaluar la muestra de validación con la sistemática *up* en el modelo de base los eventos de fondo (histograma rojo, Caso II) se clasifican peor en comparación con el caso ideal (histograma verde, Caso I). Sin embargo, cuando esta misma muestra con la sistemática *up* se evalúa en el modelo entrenado para mitigar el efecto de los errores sistemáticos que afectan a la MET, se observa una clara mejora en la clasificación de los eventos de fondo (histograma rosa, Caso III). Esto significa que, al incluir el efecto de los errores sistemáticos en el entrenamiento, el modelo resultante es capaz de discriminar mejor el fondo cuando se le pasan datos afectados por la sistemática.

Por otro lado, la subfigura de la derecha muestra la misma información, pero en escala logarítmica. Esta representación es de gran utilidad para ver el rendimiento de los diferentes modelos en torno a las colas de la distribución de los eventos de fondo, que son la zona de interés al estar luchando contra un fondo de varios órdenes de magnitud superior que la señal. En efecto, se observa que al evaluar los fondos afectados por la sistemática en el modelo de base hay varios eventos que se clasifican incorrectamente como señal (histograma rojo, en la región de *outputs* de la ANN mayores a 0.8). En particular, esto supone un empeoramiento con respecto a la situación ideal (histograma verde, Caso I), donde no había estas "falsas señales". No obstante, cuando se evalúa esta misma muestra en el modelo reducido, disminuye notablemente la proporción de eventos de fondo en la cola (histograma rosa). Precisamente, esto indica que el método reducido es capaz de mitigar el efecto del error sistemático, al reducir el número de eventos de fondo mal clasificados como señal. Esto último se traduce en un aumento efectivo del nivel de rechazo de fondo fijada la eficiencia de la señal.

4.3.3. Mitigación de los errores sistemáticos que afectan a la JES

Esta subsección tiene una estructura similar a la anterior, con la diferencia que aquí se reportan los resultados de la mitigación de los errores sistemáticos que afectan a la JES. Nuevamente, en la Figura 4.6 se comprueba que no ha habido sobreentrenamiento, mientras que en la Tabla 4.6

se muestra el poder de discriminación de las 10 variables más influyentes en la clasificación.

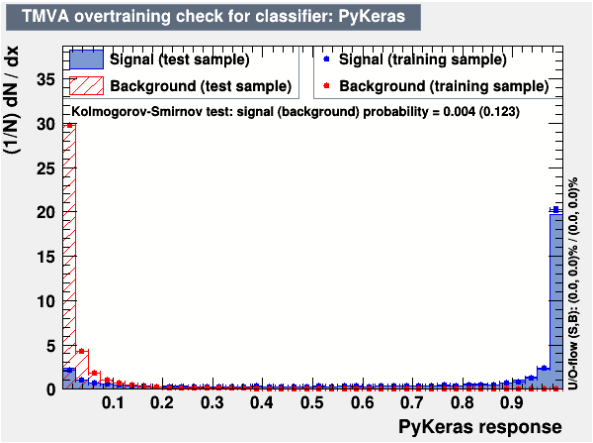

Figura 4.6: Comprobación de sobreentrenamiento del modelo reducido entrenado para mitigar los sistemáticos que afectan a la JES.

Variable	Separación
`mt2ll`	0.5936
`METcorrected_pt`	0.5004
`massT`	0.3203
`mt2bl`	0.2399
`mblt`	0.1503
`r2l4j`	0.1465
`dark_pt`	0.0257
`overlapping_factor`	0.0244
`reco_weight`	0.0243
`costhetall`	0.0233

Tabla 4.6: Poder de discriminación de las 10 variables más influyentes en el modelo reducido para mitigar los sistemáticos sobre la JES.

Una vez más, la variable `METcorrected_pt` pierde poder de discriminación, como era de esperar, pues el error sistemático que afecta a la JES se propaga a la MET.

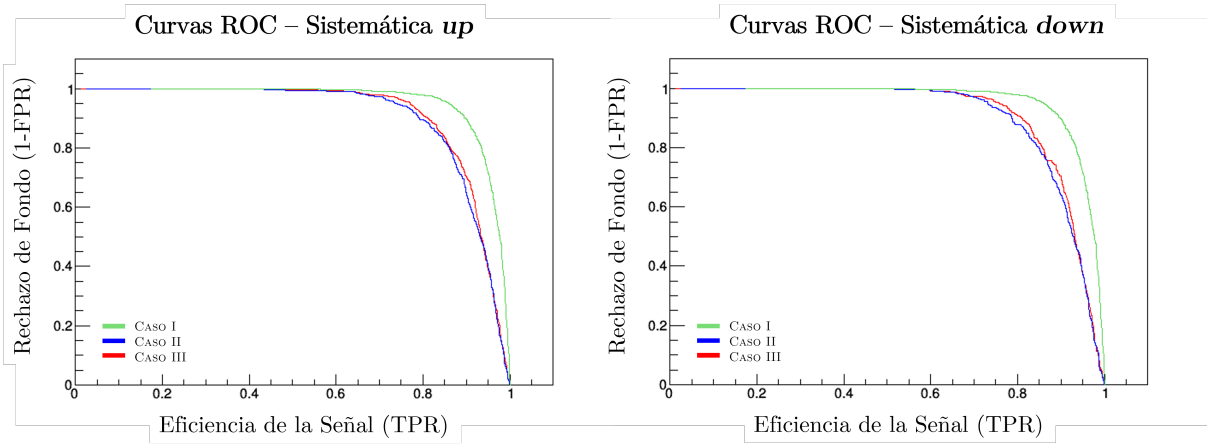

Figura 4.7: Comparativa de las curvas ROC al tratar de mitigar el efecto de los errores sistemáticos que afectan a la JES utilizando el método reducido de *data augmentation*. El significado de la leyenda es similar a la de la Figura 3.6.

En la Figura 4.7 se muestra la comparativa de curvas ROC entre el modelo de base y el modelo reducido de *data augmentation* entrenado para mitigar los errores sistemáticos que afectan a la JES. Se observa un patrón similar al obtenido al mitigar los sistemáticos que afectan a la MET, esto es, se observa que al evaluar en el modelo de base la muestra de validación con efectos sistemáticos (curva azul) el rendimiento con respecto a la situación ideal (curva verde) se degrada considerablemente. Sin embargo, al evaluar la muestra con sistemática en el modelo reducido (curva roja), se recupera algo de rendimiento. En particular, la Tabla 4.7 muestra los valores cuantitativos de la eficiencia de la señal y rechazo de fondo para algunos puntos

concretos de la ROC que resultan de interés para el análisis. La elección de estos puntos tiene la misma motivación que la dada en la sección anterior (reducir un fondo mucho más dominante, que enmascara la señal).

CASUÍSTICA	ES a 0.95 RF	ES a 0.99 RF	RF a 0.80 ES	RF a 0.95 ES	RF a 0.99 ES
Caso I	0.8591	0.7043	0.9780	0.7134	0.2835
Caso II (*up*)	0.7376	0.6289	0.8951	0.4047	0.0635
Caso III (*up*)	0.7702	0.6396	0.9129	0.4047	0.1536
Efecto Original	0.1215	0.0754	0.0829	0.3087	0.2200
Efecto Mitigado	0.0889	0.0647	0.0652	0.3087	0.1299
Ganancia (*up*)	0.0327	0.0107	0.0177	0.0000	0.0901
Reducción (%)	26.9	14.2	21.4	0.0	41.0
Caso II (*down*)	0.7317	0.6123	0.8784	0.3733	0.0497
Caso III (*down*)	0.7597	0.6274	0.9092	0.3750	0.0599
Efecto Original	0.1275	0.0920	0.0996	0.3401	0.2339
Efecto Mitigado	0.0994	0.0769	0.0688	0.3384	0.2236
Ganancia (*down*)	0.0281	0.0151	0.0308	0.0017	0.0103
Reducción (%)	22.0	16.4	30.9	0.5	4.4

Tabla 4.7: Resultados cuantitativos de la mitigación de los errores sistemáticos que afectan a la JES. Con "efecto original" nos referimos a la degradación que se produce al pasar del Caso I al Caso II, mientras que el "efecto mitigado" es la degradación del Caso I al Caso III. La ganancia es la diferencia entre el "efecto original" y el "efecto mitigado".

Entre los valores de la tabla anterior, destaca el punto en el que se tiene un 80 % de eficiencia de la señal. En este caso, el fondo mal clasificado es idealmente el 2.2 %. No obstante, al introducir el efecto de la sistemática en la JES, el porcentaje de fondo mal clasificado aumenta aproximadamente hasta el 12 %. Cuando se emplea el modelo reducido propuesto en este trabajo en lugar del modelo de base (como habitualmente se hace en LHC), se logra que el fondo mal clasificado disminuya al 9 %. Esto último supone que se tienen un 3 % más de sucesos de fondo bien clasificados con respecto al modelo de base, lo que es bastante significativo teniendo en cuenta que la sección eficaz del fondo en este problema es varios órdenes de magnitud superior a la de la señal. Análogamente, en el punto en el que la eficiencia de la señal es del 99 %, el método reducido consigue reducir en un 9 % los eventos de fondo mal clasificados. No obstante, en este último punto el fondo sigue siendo demasiado elevado, aún a pesar de la mitigación.

Finalmente, en la Figura 4.8 se muestran los histogramas con los *outputs* que resultan de evaluar las distintas muestras analizadas (nominal, con sistemática) en el modelo de base y el modelo reducido de *data augmentation*. Se muestran los resultados con la sistemática *down*, pero los obtenidos con la sistemática *up* son similares. Al igual que ocurría cuando se mitigaba la MET, el modelo reducido es capaz de clasificar los sucesos de fondo afectados por la sistemática (histograma rosa, subfigura izquierda) mucho mejor que el modelo de base (histograma rojo, subfigura izquierda), aunque sin recuperar el rendimiento del caso ideal (histograma verde, subfigura izquierda). Focalizando en las colas de las distribuciones del fondo (para *outputs* de la ANN tendiendo hacia uno), que es la región de interés para este análisis, el modelo de base clasifica una proporción elevada de sucesos de fondo como señal (histograma rojo, subfigura derecha). Esto es peligroso, porque son precisamente estos procesos de fondo mal clasificados los que pueden enmascarar la señal. Sin embargo, el modelo reducido es capaz de reducir drásticamente la proporción de fondo en esta región (histograma rosa, subfigura derecha),

mitigando así el efecto del error sistemático que afecta a la JES.

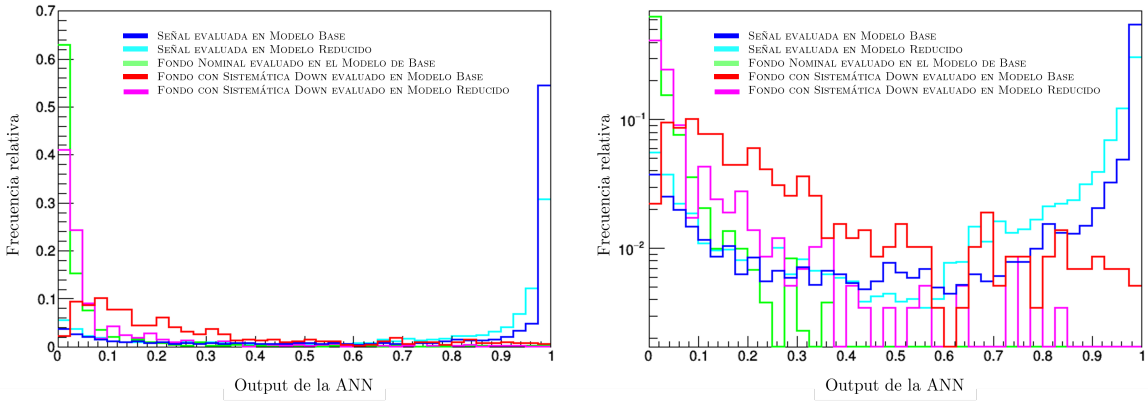

Figura 4.8: Histogramas normalizados con los *outputs* de las ANNs al tratar de mitigar el efecto sistemático que afecta a la JES. En la figura de la izquierda se representan en escala lineal, mientras que en la figura de la derecha se representan en escala logarítmica.

4.3.4. Mitigación de los errores sistemáticos que afectan a la JES y a la MET

Finalmente, se utilizó el método reducido de *data augmentation* para tratar de mitigar simultáneamente los errores sistemáticos que afectan a la JES y a la MET. Para ello, el conjunto de entrenamiento se aumentó con las muestras de entrenamiento con la sistemática asociadas tanto a la JES como a la MET. Los resultados de la comparativa de curvas ROC habitual se representan en la Figura 4.9.

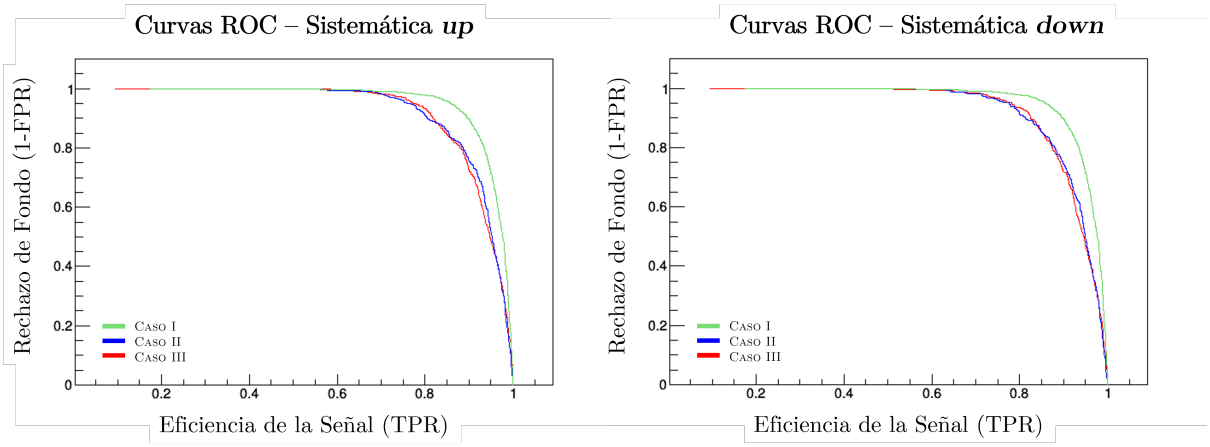

Figura 4.9: Comparativa de las curvas ROC al tratar de mitigar el efecto combinado de los errores sistemáticos que afectan a la MET y a la JES utilizando el método reducido de *data augmentation*. El significado de la leyenda es similar a la de la Figura 3.6.

Como se observa en la figura anterior, la prueba no resultó satisfactoria. La razón es que las curvas del Caso II y el Caso III son casi idénticas, lo que significa que no se recupera nada de rendimiento con respecto al modelo de base. Esto puede deberse a varios aspectos. Por un lado,

podría ocurrir que, al introducir tanta variabilidad en el conjunto de entrenamiento, el algoritmo ignore por completo las variables afectadas por la sistemática y se apoye en variables menos discriminantes para efectuar la clasificación. Una posible solución sería reducir la proporción de muestras sesgadas incluidas en la fase de entrenamiento, tratando de buscar un equilibrio entre muestras nominales y sistemáticas. Por otro lado, los hiperparámetros que optimizan las ANNs en el modelo de base, no son los mismos que optimizan el modelo reducido. Por ello, un proceso de re-optimización de los hiperparámetros que definen las ANNs podría solucionar el problema.

4.3.5. Método reducido aplicado a los BDT

En esta sección se estudia qué ocurre si se aplica el método reducido de *data augmentation* a un BDT en lugar de a una ANN. En particular, se analiza el caso en el que se trata de mitigar el error sistemático que afecta a la MET. La comparativa de curvas ROC habitual se muestra en la Figura 4.10.

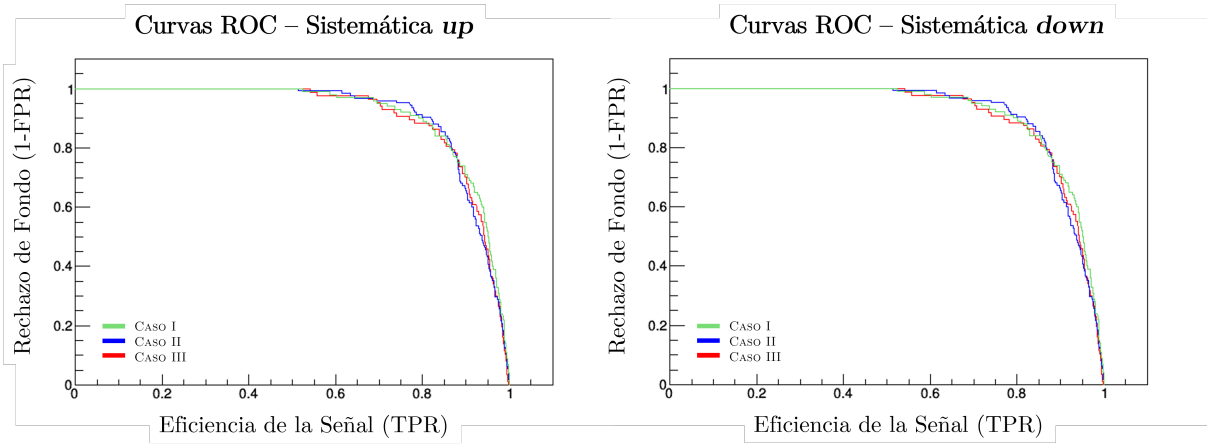

Figura 4.10: Comparativa de las curvas ROC al tratar de mitigar el efecto del error sistemático que afecta a la MET utilizando el método reducido de *data augmentation* y un BDT como método multivariante en lugar de una ANN.

Se observa que el efecto del sistemático no produce una diferencia sustancial entre el Caso I (curva verde) y el Caso II (curva azul). En particular, en ocasiones el Caso II tiene un mejor rendimiento que el Caso I. Esto quiere decir que la sistemática que afecta a la MET apenas afecta al rendimiento del BDT, haciendo que no haya margen de ganancia posible y que no tenga mucho sentido aplicar el método reducido.

En particular, al evaluar una muestra con efectos sistemáticos en un modelo entrenado exclusivamente con muestras ideales, cabría esperar una degradación en el rendimiento del clasificador. Precisamente, este empeoramiento no se observa en los BDT, lo que ha abierto una discusión en la colaboración CMS sobre si los BDT son buenas herramientas para gestionar los efectos de los errores sistemáticos.

Capítulo 5

Conclusiones

En este trabajo de fin de grado se aborda el problema de los efectos de los errores sistemáticos en tareas de clasificación con método multivariantes, especialmente orientadas al caso de análisis de física de partículas. En particular, se proponen estudian métodos para mitigar el efecto de los errores sistemáticos en la fase de entrenamiento de los métodos multivariantes.

En primer lugar, se han propuesto distintos métodos para la mitigación de la incertidumbre sistemática, todos ellos fundamentados en redes neuronales artificiales. Por un lado, se han explorado propuestas que tratan de afrontar el problema introduciendo penalizaciones en la función de pérdidas del algoritmo de aprendizaje con el objetivo de conseguir un clasificador más robusto ante observaciones afectadas por la sistemática. Por otro lado, se han estudiado técnicas de *data augmentation*, que lo que buscan es exponer al método a unos datos con mayor variabilidad con la finalidad de que los patrones de clasificación aprendidos sean menos sensibles a los errores sistemáticos que afectan a los datos. Estas propuestas se han programado y sometido a prueba en varios ejemplos sintéticos haciendo uso de las redes neuronales artificiales de la interfaz de KERAS para RStudio. Las principales conclusiones obtenidas han sido:

- Las redes neuronales permiten obtener una separación aceptable en problemas de clasificación que no son linealmente separables, como el de las esferas concéntricas expuesto en la Sección 3.1.
- El efecto de los errores sistemáticos degrada de manera significativa el rendimiento de los modelos de redes neuronales que se entrenan exclusivamente con datos nominales.
- Los métodos basados en mitigar el efecto de los errores sistemáticos mediante modificaciones en la función de pérdidas de los algoritmos de aprendizaje no proporcionan resultados positivos. Entre otras cosas, esto puede deberse a que las variables sobre las que se proyecta el efecto de los errores sistemáticos no sean continuas, a la definición escogida de los pesos o a que los pesos asignados dependan excesivamente de los datos de entrenamiento.
- Los métodos de *data augmentation* son capaces de mitigar de una manera aceptable los errores sistemáticos. En particular, el método reducido permite abordar el problema con una complejidad menor que el método tradicional, pues únicamente consiste en añadir en el conjunto de entrenamiento las muestras con la sistemática, que están a disposición del investigador en la gran mayoría de los análisis realistas de física de partículas.

Una vez explorados distintos métodos sobre ejemplos sintéticos, se aplicó el método reducido

de *data augmentation* propuesto en este trabajo a un caso realista de física de partículas, en concreto a una búsqueda de materia oscura en asociación con quarks top recientemente publicada en la tesis doctoral [46]. En particular, se ha utilizado como base el código del análisis efectuado en [46], pero se ha adaptado para poder aplicar el método reducido. Esto ha obligado a trabajar con PyROOT y con la interfaz de TMVA.

En este problema, se tiene una señal, la $t/\bar{t}$+DM, cuya sección eficaz es varios órdenes de magnitud inferior a la del fondo principal, el proceso $t/\bar{t}$(SM). Esto hace que la mitigación de los errores sistemáticos que afectan al fondo sea crucial, pues estos efectos podrían enmascarar posibles eventos de señal. En concreto, en este trabajo se han utilizado ANNs y BDTs para mitigar los errores sistemáticos que afectan a la JES y a la MET de los procesos de fondo. A continuación, se enumeran las principales conclusiones obtenidas:

- Las ANNs son un buen aliado para las tareas de discriminación entre sucesos de señal y fondo cuyas diferencias, aparentemente, no son medibles.
- El método reducido de *data augmentation* permite reducir de manera sustancial los errores sistemáticos que afectan a la MET en algunos puntos concretos de eficiencia de la señal y rechazo de fondo. En particular, es capaz de disminuir en varios puntos porcentuales la proporción de sucesos de fondo mal clasificados como señal. Esto último se traduce en una "limpieza" de las colas de las distribuciones del fondo, que es muy importante porque no hay que olvidar que estas "falsas señales" pueden enmascarar verdaderos sucesos de señal. En efecto, esto podría marcar la diferencia entre descubrir nueva física o no.
- El método también muestra un buen rendimiento al tratar de mitigar los errores sistemáticos que afectan a la JES, aunque el margen de mejora es menor.
- La mitigación del efecto combinado de varios errores sistemáticos no ha tenido los resultados esperados, pues el método reducido en ocasiones muestra una degradación con respecto al modelo de base. Una posible solución para este problema es re-optimizar los parámetros que definen la ANN, pues al introducir más datos en el conjunto de entrenamiento es de esperar que sean necesarias más neuronas para poder gestionar correctamente toda la información. Otra posible solución sería reducir el número de observaciones distorsionadas que se añaden en el entrenamiento, ya que un exceso de estas últimas podría hacer que los patrones de clasificación aprendidos se apoyen en variables quizá no tan discriminantes para el problema considerado, empeorando la clasificación.
- Se ha comprobado que al evaluar la muestra de validación con sistemática en el BDT entrenado con muestras ideales no se produce una degradación en el rendimiento. Esto hace que no exista un margen de ganancia para el sistemático, de manera que no tiene sentido aplicar el método reducido en este caso. No obstante, el uso de BDT para gestionar los errores sistemáticos es un tema de discusión actual en la colaboración CMS.
- Se espera que al añadir los efectos de los errores sistemáticos que afectan a la señal, los resultados del método reducido sean aún mejores.

En definitiva, entre los diversos métodos explorados para la mitigación del efecto de los errores sistemáticos, se ha comprobado que el método reducido de *data augmentation* es el más sencillo, el más práctico y el que mejores resultados proporciona. Esta propuesta novedosa no solo permite obtener buenos resultados en ejemplos sintéticos, sino que su efectividad también puede extrapolarse a casos realistas como el de búsqueda de materia oscura.

Apéndice A

Resultados: método de réplicas "tradicional"

En este apéndice se incluyen los resultados obtenidos al aplicar el método de réplicas "tradicional" al ejemplo de la Sección 3.1 con un error sistemático del $\pm 20\,\%$ en la variable x.

Teniendo en cuenta la descripción del método "tradicional" (véase la Sección 3.4.1), el número de réplicas de cada observación será una, puesto que se implementa un único error sistemático. En particular, cada réplica se distorsiona aleatoriamente mediante un multiplicador tomado de la distribución $\mathcal{U}$[-0.2,+0.2]. Así, si x_i es el valor de la variable x en la observación nominal i-ésima, el valor de la variable x en la réplica distorsionada será $x_i^{\mathrm{REP}} = (1+d)\,x_i$, con $d \sim \mathcal{U}$[-0.2,+0.2].

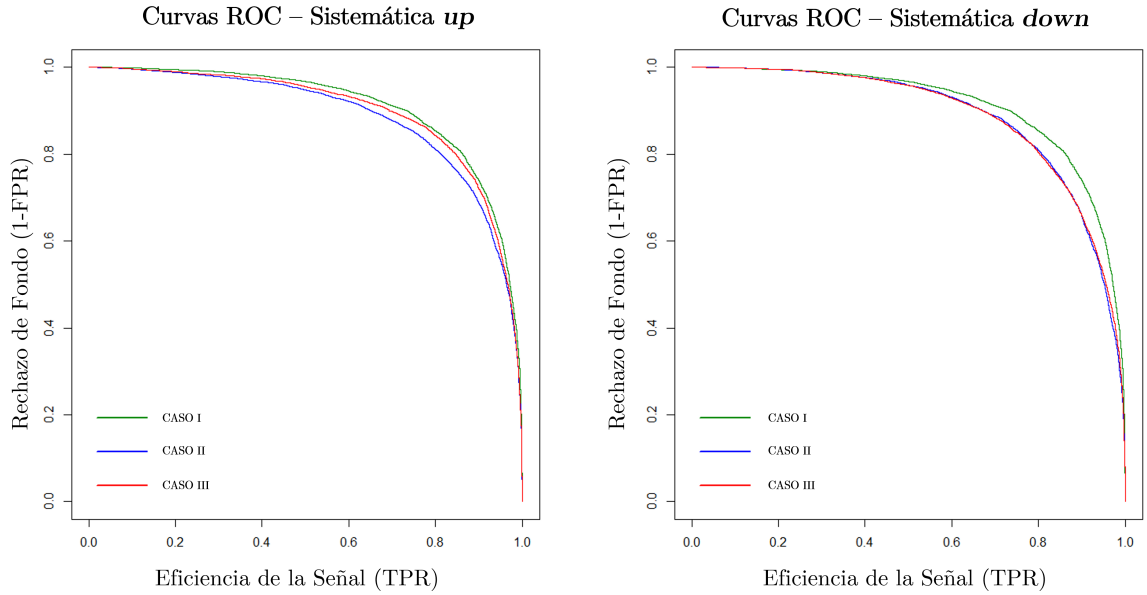

Figura A.1: Comparativa de las curvas ROC al tratar de mitigar el efecto de un error sistemático de un $\pm 20\,\%$ en la variable x mediante el método de réplicas tradicional. El significado de la leyenda es similar al de otras ocasiones (ver Figura 3.6).

La Figura A.1 muestra que el modelo de réplicas tradicional proporciona una mejora significativa con respecto al modelo de base al evaluar los datos de la muestra de validación con la sistemática *up*. En particular, el Caso III es muy próximo al Caso I, lo que significa que este método permite recuperar prácticamente el rendimiento de la situación ideal y mitigar de una manera bastante efectiva el error sistemático considerado. Por otro lado, no se observa mejora al evaluar en el modelo anterior los datos con la sistemática *down*, pero tampoco una degradación, lo cual es aceptable.

Bibliografía

[1] Pekka K Sinervo. Definition and treatment of systematic uncertainties in high energy physics and astrophysics. In *Proc. Conf. on Statistical Problems in Particle Physics, Astrophysics and Cosmology (PHYSTAT 2003), Stanford, CA, 8–11 September*, pages 122–129. Citeseer, 2003.

[2] ATLAS Collaboration. Performance of b-jet identification in the ATLAS experiment. *Journal of instrumentation*, 11(04):P04008, 2016.

[3] Albert M Sirunyan, Malte Backhaus, et al. Identification of heavy-flavour jets with the CMS detector in *pp* collisions at 13 TeV. *Journal of Instrumentation*, 13:P05011, 2018.

[4] CMS Collaboration. Search for the Higgs boson decaying to two muons in proton-proton collisions at $\sqrt{s}$= 13 TeV. *Physical review letters*, 122(2):021801, 2019.

[5] ATLAS Collaboration. Search for non-resonant Higgs boson pair production in the $bb\ell\nu\ell\nu$ final state with the ATLAS detector in *pp* collisions at $\sqrt{s}$= 13 TeV. *Physics Letters B*, 801:135145, 2020.

[6] Aishik Ghosh, Benjamin Nachman, and Daniel Whiteson. Uncertainty Aware Learning for High Energy Physics With A Cautionary Tale.

[7] Kim Albertsson, Piero Altoe, et al. Machine learning in high energy physics community white paper. In *Journal of Physics: Conference Series*, volume 1085, page 022008. IOP Publishing, 2018.

[8] Li-Gang Xia. QBDT, a new boosting decision tree method with systematical uncertainties into training for High Energy Physics. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 930:15–26, 2019.

[9] Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to pivot with adversarial networks. *Advances in neural information processing systems*, 30, 2017.

[10] Antonia Creswell, Tom White, et al. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65, 2018.

[11] Victor Estrade, Cécile Germain, Isabelle Guyon, and David Rousseau. Systematic aware learning: A case study in High Energy Physics. In *EPJ Web of Conferences*, volume 214, page 06024. EDP Sciences, 2019.

[12] Adam Elwood and Dirk Krücker. Direct optimisation of the discovery significance when training neural networks to search for new physics in particle colliders. *arXiv preprint arXiv:1806.00322*, 2018.

[13] Pablo De Castro and Tommaso Dorigo. INFERNO: Inference-aware neural optimisation. *Computer Physics Communications*, 244:170–179, 2019.

[14] Tomasso Dorigo. Systematic uncertainties: the new target of machine learning for HEP. Disponible en: URL. Accedido por última vez el 06/06/2022.

[15] CERN. The Large Hadron Collider. Disponible en: URL. Accedido por última vez el 10/06/2022.

[16] Siona Ruth Davis. Interactive Slice of the CMS detector. Technical report, 2016.

[17] CERN. The Compact Muon Solenoid. Disponible en: URL. Accedido por última vez el 10/06/2022.

[18] Stephen M Kent. Dark matter in spiral galaxies. II-Galaxies with HI rotation curves. *The Astronomical Journal*, 93:816–832, 1987.

[19] Silvia Galli, Fabio Iocco, Gianfranco Bertone, and Alessandro Melchiorri. CMB constraints on dark matter models with large annihilation cross section. *Physical Review D*, 80(2):023505, 2009.

[20] Teresa Marrodan Undagoitia and Ludwig Rauch. Dark matter direct-detection experiments. *Journal of Physics G: Nuclear and Particle Physics*, 43(1):013001, 2015.

[21] Gianfranco Bertone and David Merritt. Dark matter dynamics and indirect detection. *Modern Physics Letters A*, 20(14):1021–1036, 2005.

[22] CMS Collaboration. Search for dark matter produced in association with heavy-flavor quark pairs in proton-proton collisions at $\sqrt{s}$= 13 TeV. 2018.

[23] P. Bhat. Multivariate Analysis Methods in Particle Physics. *Annual Review of Nuclear and Particle Science*, 61:281–309, 11 2011.

[24] H.B. Prosper. Multivariate methods in particle physics: Today and tomorrow. In *XII International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT08)*, volume 3, 2008.

[25] CMS Collaboration. Observation of the diphoton decay of the Higgs boson and measurement of its properties. *The European Physical Journal C*, 74(10):1–49, 2014.

[26] ATLAS Collaboration. Measurement of Higgs boson production in the diphoton decay channel in *pp* collisions at center-of-mass energies of 7 and 8 TeV with the ATLAS detector. *Physical Review D*, 90(11):112015, 2014.

[27] A. Hoecker et al. TMVA-toolkit for multivariate data analysis, 2007. arXiv 0703039.

[28] François Chollet et al. Keras, 2015. Disponible en: URL. Accedido por última vez el 05/03/2022.

[29] Stephen Marsland. *Machine learning: an algorithmic perspective*. Chapman and Hall/CRC, 2011.

[30] Charu C Aggarwal et al. Neural networks and deep learning. *Springer*, 10:978–3, 2018.

[31] Jose Manuel Gutiérrez et al. *Redes Neuronales y Probabilísticas en las Ciencias Atmosféricas*. Monografías del Instituto Nacional de Meteorología. Ministerio de Medio Ambiente, Madrid, 2004.

[32] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[33] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151, 1999.

[34] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[35] Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012.

[36] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6980, 2015.

[37] ATLAS Collaboration. Jet calibration and systematic uncertainties for jets reconstructed in the ATLAS detector at $\sqrt{s}$=13 TeV. Technical report, ATL-PHYS-PUB-2015-015, 2015.

[38] ATLAS Collaboration. Performance of missing transverse momentum reconstruction with the ATLAS detector using proton-proton collisions at $\sqrt{s}$= 13 TeV. *arXiv preprint arXiv:1802.08168*, 2018.

[39] Niall McHugh. Jet Energy Resolution and Scale Measurements for the ILD using Durham and MC Jet Clustering with $Z \longmapsto qq$ Events. 2018.

[40] CMS Collaboration. Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV. 2017.

[41] Tomasz Kalinowski and François Chollet. R interface to Keras. Disponible en: URL. Accedido por última vez el 05/03/2022.

[42] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[43] Louis Lyons. A method of reducing systematic errors in classification problems. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 324(3):565–568, 1993.

[44] Luis Crespo Ruiz. Aplicación a física de partículas de métodos de clasificación multidimensionales en presencia de errores sistemáticos. Universidad de Cantabria. TFG.

[45] Sylvain Fichet. Taming systematic uncertainties at the LHC with the central limit theorem. *Nuclear Physics B*, 911:623–637, 2016.

[46] Cedric Prieels. Search for dark matter production in association with top quarks in the dilepton final state at $\sqrt{s}$=13 TeV. Universidad de Cantabria. Tesis doctoral.