Comparativa de algoritmos de clasificación multivariable (MVA)

Trabajo de fin de Grado de Física

Óscar Prieto Amigo

Tutores:
Francisco Javier Cuevas Maestro
Carlos Francisco Erice Cid

Facultad de Ciencias, Oviedo. Curso 2019 - 2020.

Índice

1. Introducción

El objetivo principal de este este trabajo es el estudio de un tipo de proceso de *nueva física*. Para esto introduciremos el **Modelo Estándar**, una teoría que explica un gran rango de fenómenos naturales pero que dista de ser una descripción completa de la naturaleza. Veremos como para explicar fenómenos que no son descritos por este hemos de recurrir a posibles extensiones del modelo. En nuestro caso el estado final del proceso estudiado tiene partículas que no están descritas por el Modelo Estándar (son candidatas a materia oscura); para solucionar esto podemos valernos de una de las teorías que explica este fenómeno y es consistente con el Modelo Estandar: la supersimetría.

Para el estudio de estos fenómenos necesitamos un montaje específico para la toma de datos; explicaremos detenidamente las características del **LHC**, el detector **CMS** y sus componentes, así como los procesos de identificación y reconstrucción de evento que se llevan a cabo para facilitarnos estos datos sobre los que realizar el análisis.

Después de la toma de datos profundizaremos en el proceso supersimétrico estudiado, definiendo la señal y el fondo, y haciendo uso de simulaciones estableceremos cuáles son nuestras regiones de interés y por qué lo son. Veremos que en ciertas regiones gran parte del fondo es debido a un error al establecer de dónde proceden ciertas partículas del estado final; cabe entonces la posibilidad del desarrollo de un algoritmo clasificador que ayude en esta tarea y minimice en la medida de lo posible este error para reducir así el fondo y con ello mejorar nuestra medida.

El paso siguiente será el estudio de un método multivariable que establezca un nuevo etiquetado con una menor tasa de error que el etiquetado de partículas original. En nuestro caso optaremos por una red neuronal como clasificador, por su facilidad de implementación y modulado.

Finalmente aplicaremos este nuevo etiquetado y compararemos en todas las regiones de interés las predicciones respecto al etiquetado original para evaluar la mejoría en la medida cuantitativamente.

2. Modelo Estándar y extensiones

2.1. Modelo Estándar

Se llama Modelo Estándar (que abreviamos como SM, de *Standard Model*) a la teoría que describe el conjunto de interacciones fuerte, débil y electromagnética; sirve para caracterizar estas interacciones y las partículas que las sufren o median.

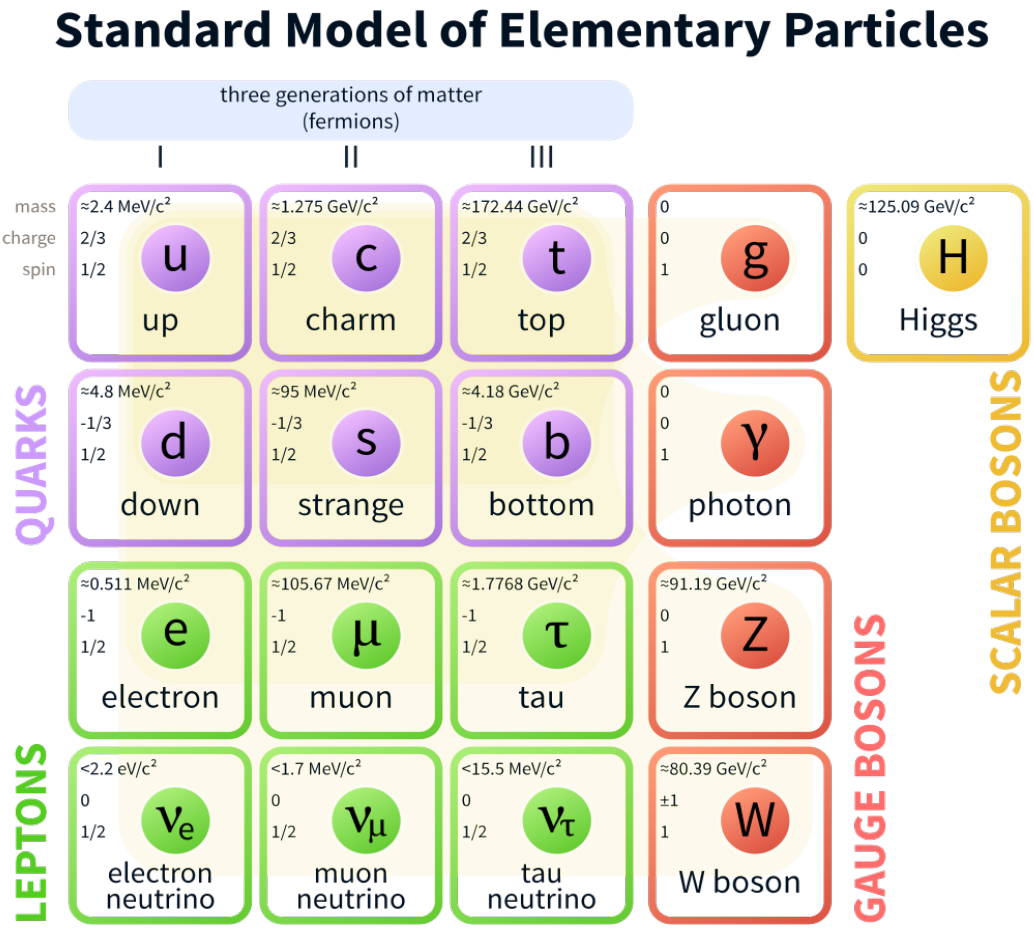

Figura 1: Partículas del SM

Estas partículas, que podemos ver en la figura 1, se dividen en dos grupos atendiendo a su espín. Las partículas de espín 1/2 son los fermiones, constituyentes de la materia, que siguen la estadística de Fermi-Dirac y cumplen el principio de exclusión de Pauli. Dentro de esta familia se puede hacer una subdivisión dependiendo de si las partículas sufren interacción fuerte; si este es el caso nos referimos a los quarks, de no ser así (partículas que

solo se relacionan a través de la interacción electrodébil) hablamos de leptones.
Las partículas de espín entero son los bosones, que median las interacciones fundamentales y siguen la estadística de Bose-Einstein. En el SM hay 5 bosones:

- Bosón de Higgs, bosón escalar que explica la masa de la mayor parte de las partículas fundamentales
- Gluón, que media la interacción fuerte
- Fotón, un mediador de la interacción electromagnética
- Z, mediador de la interacción débil con carga eléctrica nula
- $W^{\pm}$, mediador de la interacción débil con carga eléctrica $\pm e$

El SM es una teoría sólida que explica y predice un gran rango de los fenómenos que actualmente nos es posible conocer, pero hay algunos fenómenos que no son objetivo de su descripción, obligándonos a recurrir a posibles extensiones del modelo en estos casos.

2.2. Necesidad de la ampliación del SM

Uno de los hechos más llamativos en lo referente al SM es la no incorporación de una de las fuerzas fundamentales, la gravedad. Esta interacción viene descrita por la relatividad general (Einstein, 1915). Por suerte y desgracia, la teoría cuántica de campos (QFT, base del SM) y la relatividad general funcionan en su rango de aplicabilidad (masas y distancias pequeñas, masas y distancias grandes respectivamente), pero tal como están formuladas en la actualidad no es posible una unificación de ambas. Esto es, si nos movemos en una escala energética similar a la masa de Planck reducida ($2{,}4 \times 10^{18}$ GeV), los efectos de la gravedad cuántica no pueden ser despreciados y por tanto las predicciones del SM carecen de validez.
Si compraramos la escala de Plank reducida con el orden energético del sector electrodébil (digamos la masas de los bosones W y Z, 100 GeV) vemos una diferencia de 16 órdenes de magnitud. ¿Qué causa este problema de jerarquía, este contraste entre masas necesarias para observar los efectos de estas diferentes interacciones?

Otro de los temas que se escapa a una descripción completa en el marco del SM es la asimetría materia-antimateria presente en el universo. Esto podría venir explicado en parte por una violación de la simetría CP, permitida por el SM y probada experimentalmente (como en el caso del proceso μ a 3 electrones). La "C" representa la conjugación de carga, esto es, la sustitución de la partícula por su correspondiente antipartícula, y la "P" hace referencia a la paridad espacial con el correspondiente cambio de espín. La violación de la simetría CP es

necesaria, pero para que la asimetría bariónica ocurriera también debería haberse dado el caso de una violación del número bariónico, y esto va más allá del SM.

Este preámbulo finaliza con una vieja conocida de las otras muchas incógnitas, la materia oscura. Ya desde 1884 Kelvin observó que la masa total de la galaxia no era compatible con la suma de las estrellas contenida por ella y concluyó que debían de existir cuerpos celestes que no eramos capaces de observar. Posteriormente ,en 1933, Zwicky aplicó el teorema del virial a un cúmulo globular de galaxias y llegó a la misma conclusión que Kelvin, había un cierto tipo de materia oscura (la llamaremos DM, de *Dark Matter*) que hacía que la masa total del cúmulo globular de galaxias fuera unas 400 veces superior a la suma de las galaxias observables. Esto es, la DM interacciona gravitacionalmente, pero no electromagnéticamente, de ahí que no la podamos observar. ¿Qué explicación a esta materia nos puede aportar el SM? Ya que no interacciona electromagnéticamente, deberían de ser los neutrinos, pero estos no tienen la suficiente masa como para ser los responsables. En el siguiente apartado veremos como podría ayudar una extensión supersimétrica al SM, dando lugar a nuevos candidatos a materia oscura.

2.3. SUSY y MSSM

2.3.1. Supersimetría

La supersimetría (SUSY) es una posible extensión a las simetrías clásicas de las teorías de campos que introduce una relación entre bosones y fermiones: predice que cada una de estas partículas tiene un supercompañero cuyo espín difiere de la partícula en 1/2. El supercompañero de un bosón será un fermión, y viceversa.
Con esta diferencia de espín se introducen otras características. Por ejemplo, un fermión tiene que cumplir el principio de exclusión de Pauli, pero su bosón asociado no tiene esta restricción.
Los compañeros escalares de los fermiones serán los sleptones y squarks, los compañeros fermiónicos del Higgs serán higgsinos y los de los bosones, gauginos.

2.3.2. Extensión supersimétrica mínima del Modelo Estándar

El MSSM (*Minimal Supersymmetric Standard Model*) es la mínima extensión supersimétrica del Modelo Estándar. No hemos observado compañeras supersimétricas en ningún experimento realizado hasta el momento, por lo que no puede ser de por sí una simetría de la naturaleza, esto es, ha de estar rota[1].

[1]En caso de que fuera una simetría perfecta, las supercompañeras tendrían la misma masa que sus compañeras del Modelo Estándar

El MSSM asigna a cada fermión un campo escalar complejo con dos compañeras supersimétricas llamadas sfermiones. Los fotones y gluones tienen otras dos compañeras, fotino y gluino. Los bosones electrodébiles tendrán sbosones llamados winos y binos y por último el bosón de Higgs tendrá los higgsinos. Los estados masivos neutros mezcla de gauginos y higgsinos se conocen como neutralinos, y si son masivos y cargados se conocen como charginos. Los neutralinos son fermiones de Majorana (son su propia antipartícula) y los charginos son fermiones de Dirac (diferentes a su antipartícula).

3. Detección y procesamiento de datos

3.1. LHC

El LHC es el mayor acelerador de partículas construido hasta el momento. Está situado en la frontera Suiza-Francia, en un túnel a 175 m bajo tierra, de 27 kilómetros de circunferencia, y su principal objetivo es la colisión de protones o iones pesados en puntos de detección; las colisiones de protones tienen en general una energía en centro de masas de 13 TeV y una luminosidad típica de $10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$.

La alta energía de colisión se consigue acelerando dos haces de partículas cargadas en sentidos opuestos mediante imanes superconductores cuadropolares que producen enormes campos electromagnéticos oscilantes; se ejerce una fuerza sobre las partículas en su sentido de movimiento en todo momento. Esta aceleración constante permite contrarrestar en gran medida las pérdidas energéticas, como por ejemplo la radiación sincrotrón, cuya existencia es consecuencia directa de la forma circular de nuestro acelerador, o colisiones con partículas residuales existentes dentro del vacío. Además es necesaria la presencia de imanes multipolares de orden superior para mantener la coherencia del haz y estabilizar su forma.

El proceso de aceleración es progresivo y comienza en un acelerador lineal (LINAC 2 o 3, dependiendo del tipo de partícula acelerada), del que las partículas pasan a un pequeño acelerador circular (*Proton Synchrotron*) y posteriormente a otro acelerador circular con mayor circunferencia (*Super Proton Synchrotron*); en estos aceleradores circulares es donde las partículas se separan en haces que circulan en sentidos contrarios. Finalmente, al alcanzar la energía deseada estos son introducidos al LHC donde se continua con el proceso de aceleración/colisión. En la figura 2 se muestra la distribución espacial de estos aceleradores y el resto de estructuras componen el LHC.

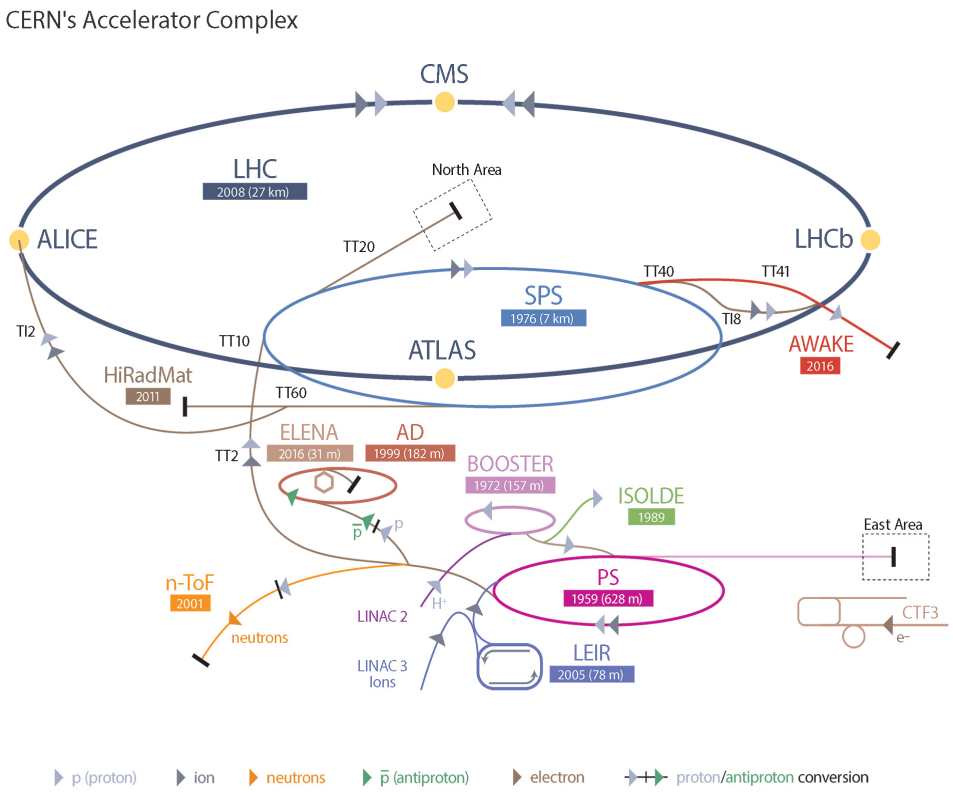

Figura 2: Distribución espacial del LHC

El LHC cuenta con cuatro detectores principales, dos de cáracter general (CMS y ATLAS) y dos de propósito específico (ALICE y LHCb).

3.2. CMS

El detector CMS (*Compact Muon Solenoid*), cuya estructura podemos ver en la figura 3, es uno de los cuatro detectores principales del LHC. Es de propósito general y está diseñado para medir con precisión las propiedades de leptones, fotones y productos hadrónicos tanto en colisiones protón-protón como en colisiones de iones pesados. Tiene unas dimensiones de 15 m de diámetro y 21 m de largo, y un peso de cerca de 14000 toneladas.

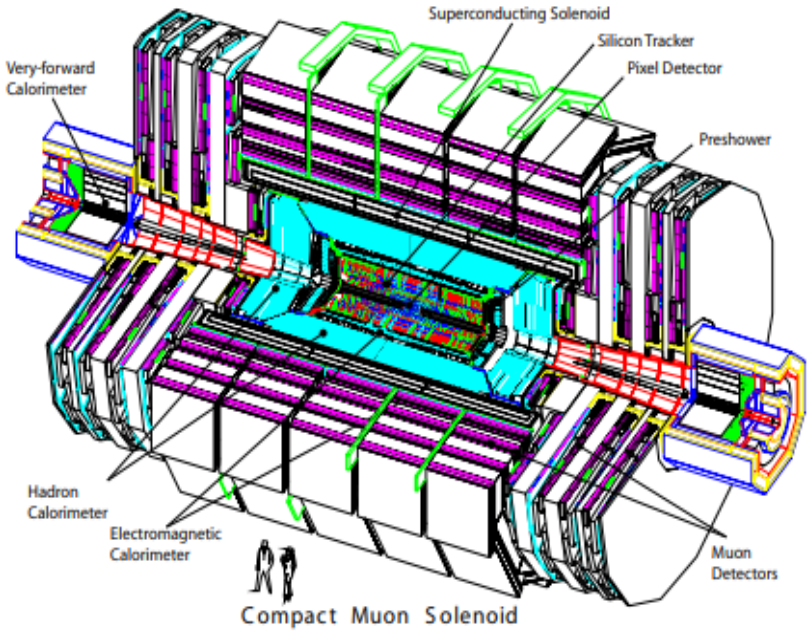

Figura 3: Vista en perspectiva del CMS

Las principales características distintivas del detector son el uso de un solenoide central superconductor, un calorímetro electromagnético basado en cristales centelleadores y un detector de trazas ultradenso de silicio.

El sistema de coordenadas que usaremos será el siguiente: consideramos el eje principal, por el que circulan los haces de partículas, como el eje Z. Los ejes X e Y forman el denominado **plano transverso**, estando el eje X en horizontal y el eje Y en vertical. Definimos a partir de este sistema (x, y, z) las coordenadas esféricas (r, θ, ϕ) asociadas. r es la distancia al origen de coordenadas, ϕ irá desde 0 hasta 2π y θ irá de $-\pi/2$ a $\pi/2$. A partir de estas últimas podemos definir la **pseudorapidez** como $\eta = -\log(\tan(\theta/2))$ y también definimos ρ como la proyección de nuestro punto sobre el plano XY, de esta manera tenemos el sistema (ρ, η, ϕ).

3.2.1. Detector de trazas

El sistema de detección de trazas, representado en la figura 4, envuelve al punto de interacción en una longitud de 5.8 m y un diámetro de 2.5 m. Está diseñado para medir precisa y eficientemente las trayectorias de las partículas eléctricamente cargadas que emergen de las colisiones con una aceptancia de $\mid \eta \mid < 2.5$ además de permitir una reconstrucción precisa de los vértices de interacción secundarios.

En promedio, a una luminosidad instantánea típica, por cada interacción de paquetes de protones habrá 20 interacciones sobrepuestas en el mismo evento, de las cuales resultan 1000 partículas. Como habrá uno de estos cruces de paquetes cada 25 ns, se necesita, por tanto, un detector con una granularidad y un tiempo de respuesta

lo suficientemente bajo, de tal manera que podamos asignar las interacciones al evento correcto. Para satisfacer estas necesidades nos encontramos con ciertos problemas.

El primero es que necesitamos mucha energía para la electrónica asociada a este detector, y para paliar el calentamiento necesitaremos sistemas de refrigeración, de modo que se incrementa el material usado y aumenta así la probabilidad de *scattering*, interacciones nucleares, radiación de frenado y producción de pares, procesos que alteran las partículas producidas en la colisión empeorando nuestras medidas. Otro problema al que nos enfrentamos es el daño por radiación que es susceptible a sufrir el detector de trazas, ya que el flujo de partículas es muy alto. La tecnología de silicio usada en el detector de trazas (característica del CMS) permite minimizar estos inconvenientes.

El detector de trazas está formado por dos tipos de detector: píxeles y tiras.

- El detector de píxeles es el más cercano a la región de interacción y abarca una pseudorrapidez de $\mid \eta \mid < 2.5$. Está formado por cuatro partes cilíndricas y dos bases con hueco central para estos, con 48 millones y 18 millones de píxeles cada una respectivamente. Cada uno de estos píxeles tiene unas dimensiones de $100 \times 150\,\mu m^2$.

  Su principal función es hacer posible la reconstrucción de vertices secundarios correspondientes a desintegraciones de b y τ.

- El detector de tiras está formado por sensores tipo microtira p-n. Está subdividido en 4 partes, como se indica en la figura. En cada parte, las tiras tienen diferentes tipos de geometría de tal manera que se cubra la superficie total. El número total de sensores asciende a 24244, ocupando un área de casi $200\,m^2$.

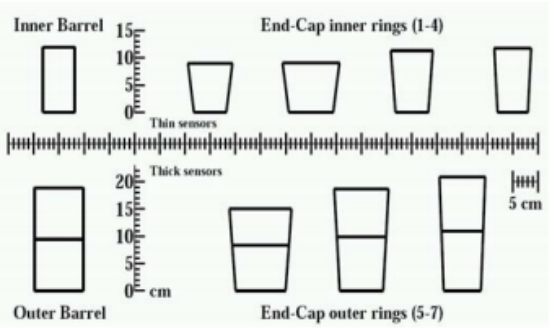

Figura 4: Vista del detector de trazas [4]

El funcionamiento del detector de trazas es el siguiente: el detector de trazas está subdividido en secciones pequeñas, como hemos visto. Las partículas cargadas eléctricamente interaccionan con electrones del silicio

y forman pares electrón-hueco que son detectables por los sistemas electrónicos acoplados y se analizan en conjunto, detectándose así deposiciones energéticas debidas a partículas.

3.2.2. Calorímetro electromagnético (ECAL)

El enorme y homogéneo calorímetro electromagnético del CMS se encuentra envolviendo el detector de trazas. Está formado por 61200 cristales de wolframato de plomo (II) ($PbWO_4$) colocados en el cilindro central como se muestra en la figura 5, y este casi cerrado por 7324 cristales *endcap* que forman tapas con hueco central frente a las cuales se encuentran detectores.

Su objetivo principal es la medida de la energía de partículas con carga electromagnética. Esto se consigue mediante centelleado: las partículas cargadas interaccionan con los electrones del material, excitándolos; al volver los electrones al estado energético fundamental se emiten fotones. Es necesario el uso de fotomultiplicadores para amplificar la señal, ya que es de baja intensidad; teniendo en cuenta la amplificación de la señal se determina la energía que la partícula ha depositado en su interacción con el material del calorímetro.

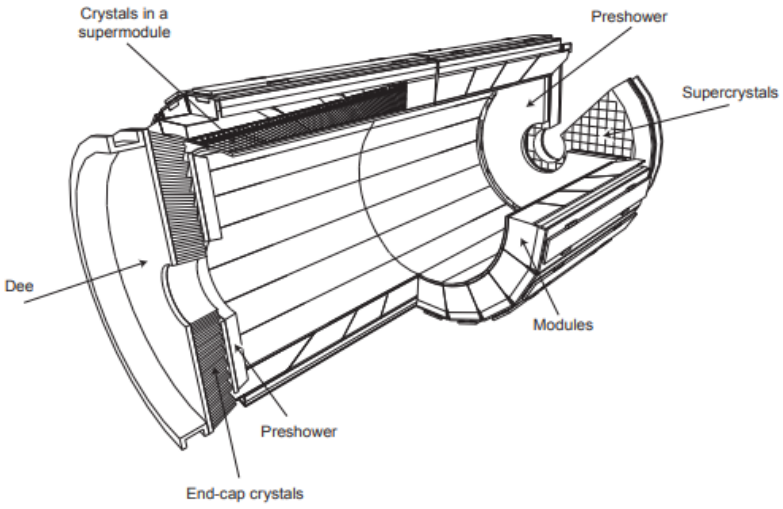

Figura 5: Vista del ECAL [4]

La parte cilíndrica del ECAL abarca una pseudorrapidez de $|\eta| < 1.479$, mientras que los *endcaps* cubren el rango $1.479 < |\eta| < 3$.

3.2.3. Calorímetro hadrónico (HCAL)

En la capa superior al ECAL se encuentra el calorímetro hadrónico, cuyas funciones incluyen la medida de *jets*[2] y contribuir a la determinación del p_T^{miss} [3]).

Su diseño y funcionamiento es similar al del ECAL, usando también el centelleo como base para la detección de partículas de carga eléctrica neutra. Para esto se usan capas de materiales muy densos (latón, entre otros) que eleven la probabilidad de interacción fuerte entre la partícula neutra y el material, dando lugar a cascadas de partículas que se amplían al atravesar las consecutivas capas y hacen posible su detección.

3.2.4. Solenoide superconductor

El solenoide superconductor, con una longitud de 12.5 m y un diámetro de 6 m, formado por una aleación de NbTi y situado entre el calorímetro hadrónico y las cámaras de muones consigue un campo en el interior del detector de cerca de 4 T. Este campo curva las partículas eléctricamente cargadas, permitiendo la rápida identificación de su carga así como el momento transverso de las partículas.

3.2.5. Sistema de detección de muones

El sistema de detección de muones está situado por encima del solenoide superconductor, en la parte más externa del detector CMS. Como da a entender el nombre del detector (*Compact Muon Solenoid*), el diseño de este viene moldeado por el requerimiento de una detección de muones precisa, confiable. Es por esta necesidad que el sistema de detección de muones ha sido diseñado para cubrir todo el rango de aceptancias, $\mid \eta \mid < 2.5$, sin ningún tipo de hueco e incluso habiendo solapamiento[4] entre partes del sistema.

El sistema general se divide en tres tipos de detectores gaseosos:

- **Cámaras de tubos de deriva (*Drift Tube System*, DTS)**

  Se localizan en las capas cilíndricas (parte *barrel*). Cada cámara contiene 2 o 3 supercapas formadas por 4 capas de celdas de deriva escalonadas de menor a mayor. Dentro de una capa hay aproximadamente 60 tubos; cada tubo de 4 cm contiene gas y un hilo tenso cargado positivamente, estando cargada negativamente la superficie del tubo. Cuando una partícula ionizante atraviesa el gas, los electrones liberados del gas son atraídos por el hilo y detectados.

  Cada supercapa proporciona gran ayuda en la identificación de los muones así como en su etiquetado temporal, con resolución del orden de nanosegundos.

[2] Haz de partículas emitido en una cierta dirección, producto de la hadronización de quarks y gluones.
[3] Energía que se espera tras un evento por conservación de la energía y el momento pero que no es detectada.
[4] Estos solapamientos presentan una región con caída no homogénea en $0{,}8 < \eta < 1{,}2$.

- **Cámaras de tiras catódicas (*Cathode Strip Chambers*, CSC)**

  Se localizan en los discos *endcap*. Cada cámara consiste en un volumen de gas en el que en una dirección del plano hay tiras de carga negativa, y en la otra dirección hay hilos de carga positiva. Al pasar un muón se ionizan electrones del gas que van a los cables, creando avalanchas de electrones. Además los iones positivos se ven repelidos por el hilo y van hacia un cátodo de cobre, creando así un pulso en las tiras. Como tenemos dos direcciones del espacio cubiertas nos proporciona dos coordenadas de posición para la partícula.

  Las CSC son bastante flexibles en cuanto a condiciones experimentales, ya que proporcionan medidas muy precisas bajo campos magnéticos altos y no uniformes.

- **Cámaras de placas resistivas (*Resistive Plate Chambers*, RPC)**

  Se encuentran tanto en la parte cilíndrica como en los discos *endcap*. Las RPC consisten en dos placas paralelas de cargas diferentes, de material altamente resistivo y separadas por un volumen de gas. Funcionan como los sistemas previamente explicados, valiéndose de las corrientes de avalancha.

  La RPC tienen buena resolución temporal (1 ns), facilitando el etiquetado correcto de muones en gran medida.

Estos detectores además contribuyen al sistema de selección de eventos *online* del detector CMS (llamado *trigger*) por su rápida respuesta.

3.2.6. *Trigger*

A una luminosidad de $10^{34}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$, CMS detectaría unos 10^9 eventos por segundo, de los que la gran mayoría carecen de relevancia en las búsquedas de nueva física. Para reducir el número de eventos a analizar el detector CMS usa un sistema de selección de eventos *online*, llamado comúnmente *trigger*, que cuenta con dos niveles:

- **L1 *trigger***, un sistema de hardware que actúa durante la detección con latencia fija (4 μs) y que utiliza información de calorímetros y detectores de muones para hacer una primera selección de los eventos que podrían ser relevantes, es decir, que contengan combinaciones inusuales de partículas o altas energías.

- ***High level trigger* (HLT)**, un conjunto de procesadores que filtran los procesos que han pasado la preselección establecida por el L1 usando toda la información disponible en el detector. Este segundo trigger actúa de manera similar a los algoritmos de reconstrucción de evento que se usan de manera *offline*, ya que no es necesario que sea tan rápido como el L1.

Gracias al *trigger* el número de eventos almacenados para su posterior análisis se reduce a unos 1000 por segundo para luminosidades instantáneas de trabajo típicas.

3.3. *Reconstrucción de partículas en CMS*

El algoritmo de reconstrucción de evento usado en CMS se denomina *particle flow* (PF). La idea fundamental es identificar todas las partículas estables resultantes del proceso (electrones, muones, fotones y hadrones) y propocionar una medida lo más fiable posible de su energía, momento y de características adicionales del proceso a través de una combinación de la información proporcionada por todos los subdetectores del sistema.

Una partícula hace aparecer varios elementos PF en los subdetectores (se consideran elementos PF a las trayectorias reconstruídas de las partículas cargadas, los vértices, las aglomeraciones en los calorímetros...). La reconstrucción de esta partícula comienza con un algoritmo que nos permite conectar diversos elementos PF de varios subdetectores: el *link algorithm*.

A grandes rasgos, la asociacion de elementos PF sigue un orden secuencial, dependiendo del tipo de partícula al que pueden estar asociados:

1. **Muones**, usando la información del detector de trazas y el sistema de detección de muones, dejan poca energía en los calorímetros.

2. **Electrones y fotones aislados** (han de reconstruirse al unísono por la existencia de radiación de frenado y de producción de pares debido a la densidad del detector de trazas), usando la información del ECAL y del detector de trazas para electrones.

3. **Hadrones y fotones no aislados**. Los quarks y gluones se producen en grandes cantidades e inmediatamente tras la colisión se agrupan en conjuntos de carga de color neutra (se **hadronizan**), por lo que se obsevarán *jets*. Los hadrones neutros y fotones no aislados se han de identificar a la vez ya que no dejan trazas, y los hadrones cargados estarán vinculados a una energía calorimétrica compatible con una traza y su momento asociado.

   Una parte importante de la caracterización de *jets* es la asignación de la probabilidad de que provengan de un quark b, que se denomina *b-tagging*. Este etiquetado es de vital importacia ya que en las colisiones se produce una gran cantidad de quarks b, y si no son nuestro objeto de estudio (como es nuestro caso) los productos de sus desintegraciones constituirán una fracción significativa del fondo que podremos eliminar estableciendo los correspondientes cortes.

4. **Partículas secundarias**, que denominamos así no por su falta de importancia si no por su procedencia, ya que vienen de interacciones que no tienen lugar en el eje principal.

5. **Partículas no detectables**, por ejemplo neutrinos y LSP, que no interaccionan con los subdetectores. Sabemos de su existencia por la presencia de **energía transversa faltante**, que será la suma de los momentos transversos de estas partículas que no detectamos.

Los elementos PF identificados en cada parte se "retiran" de tal manera que no vuelven a ser procesados.

Las ventajas que conseguimos usando un algoritmo de este tipo son varias:

- Mejora la reconstrucción de todos los objetos físicos, llegando incluso al doble de resolución para algunos, como *jets* y p_T^{miss}
- El análisis de datos se basa prácticamente solo en partículas de alta vida media, dando mucha estabilidad
- Su desempeño en datos reales es similar al que tiene sobre simulaciones, permitiendonos probarlo o modificarlo con fuidez

4. Procesos y regiones de interés

Todos los datos y resultados de búsquedas previas usados en los siguientes apartados se corresponden con colisiones pp a $\sqrt{s} = 13$ TeV y a una luminosidad integrada de 35.9 fb^{-1} recabados en el detector CMS del LHC durante la toma de datos del año 2016.

4.1. Señal

El proceso supersimétrico estudiado que consideramos señal es la producción electrodébil de charginos $\tilde{\chi}_1^{\pm}$ y neutralinos $\tilde{\chi}_2^0$, su desintegración a $\tilde{\chi}_1^0$ (LSP) y bosones W o Z repectivamente y la posterior desintegración leptónica de los W y Z proporcionando estados finales de 3 leptones y un p_T^{miss} relativamente grande (debido al ν proveniente de la desintegración del W y a 2 $\tilde{\chi}_1^0$). Estos procesos son esquematizados en las figuras 6, 7 y 8, y los denominamos TChiWZ; estudiaremos tres posibles configuraciones de las masas de $\tilde{\chi}_1^{\pm}$, $\tilde{\chi}_2^0$ y $\tilde{\chi}_1^0$ ya que son parámetros desconocidos.

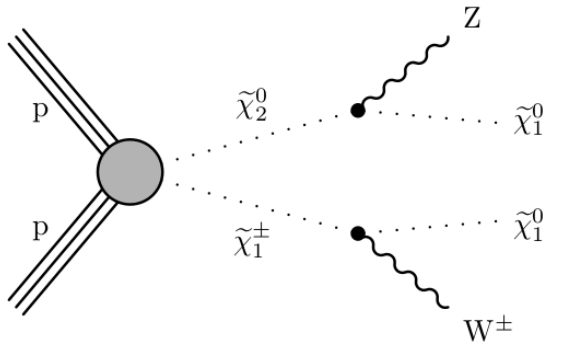

Figura 6: Proceso TChiNeuWZ [1]

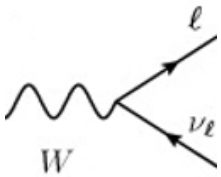

Figura 7: Desintegración leptónica de W

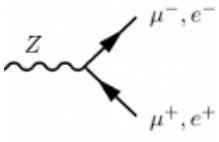

Figura 8: Desintegración leptónica de Z

Concretando aún más, nos interesan los procesos en los que estos leptones del estado final son ligeros (los 3 son e o μ, más fáciles de reconstruir en CMS); entre estos habrá dos que provengan del bosón Z y por tanto que formen un par de leptones con signos opuestos y mismo sabor, comúnmente denominado "par OSSF"(*Opposite Sign Same Flavor*). No conocemos con certeza el origen de estos leptones, por lo que la asignación de su procedencia puede ser en algunos casos equívoca, como veremos más adelante.

4.2. Fondo

Se denomina fondo a todos los procesos que no son objetivo de estudio pero que tienen estados finales parecidos a la señal, por lo que son fuentes de incertidumbre en nuestra medida. Los procesos SM que podemos considerar fondo pertenecen a las siguientes categorías:

- Fondo WZ: producción de WZ o Wγ^*. Este es el fondo principal y es fundamental en este estudio, debido a que cuando los W y Z/γ^* se desintegran leptónicamente se producen estados finales de 3 e o μ (con par OSSF) y una cantidad significativa de p_T^{miss} debida al neutrino producido en la desintegración del W.
- Identificación errónea de leptones e o μ que no son producidos mediante el proceso señal, si no por procesos $t\bar{t}$, DY o W+jets. Este fondo es evitable en gran medida gracias al la selección *tight* (criterios de selección de leptones de alta calidad), a establecer cortes cinemáticos y a rechazar algunas características de evento (como la presencia de jets etiquetados como provenientes de un quark b para evitar el fondo $t\bar{t}$, por ejemplo).
- Conversiones asimétricas de fotones que acompañan a la producción del W o el Z en las que uno de los leptones creados tiene un p_T pequeño y puede no entrar en los criterios de selección.
- Procesos poco probables, como son la producción de múltiples bosones o la producción de bosón más un par $t\bar{t}$, que pueden dar estados finales similares al estudiado. La contribución total de estos procesos raros se estima mediante simulaciones Monte Carlo.

4.3. Simulaciones Monte Carlo

En física de altas energías las simulaciones Monte Carlo nos ayudan en la predicción e interpretación de datos recogidos en los detectores gracias a una reproducción precisa de los eventos generados y de la interacción de los productos de estos eventos con los detectores.

Para la simulación de evento, que nos permite conocer la sección eficaz de un proceso, se divide el proceso en varias partes:

1. Se empieza la simulación calculando la distribución de probabilidad de la mayor transferencia de momento entre partones[5]. Se considera entonces que los partones que interaccionan para que ocurra esta transferencia son los que sufren *scattering*. También se simulan los residuos de la colisión.

[5]Partículas constituyentes de los hadrones, es decir, quarks y gluones.

2. Se simula la cascada de partones producida por las partículas con carga de color aceleradas en el proceso de *scattering* o creadas/aniquiladas en procesos de producción/aniquilación partícula-antipartícula.

3. Se simulan los procesos de confinamiento de los partones en hadrones (hadronización).

4. Se simulan la evolución, hadronización y potenciales interacciones entre los residuos hadrónicos del primer paso.

5. Se simulan las posibles desintegraciones de partículas que pueden llegar a los detectores.

Respecto a la parte de detección, se simula el paso de las partículas por los materiales del detector y la respuesta de este. En general estas simulaciones de detección se dividen en dos tipos:

- Simulación completa: usada para el análisis final, ya que proporciona una aproximación fiel a la realidad pero a costa de necesitar una potencia de computación inmensa (en su paso por el material, las partículas interaccionan con este y crean una ingente cantidad de partículas secundarias que han de ser consideradas).

- Simulación rápida: se utiliza en las etapas preliminares del análisis y cuando la simulación completa es demasiado lenta debido a un espacio de fases grande. Usa simplificaciones que mejoran el tiempo de computación pero no es tan precisa como una simulación completa.

4.4. Regiones cinemáticas de interés

Gracias a simulaciones Monte Carlo podemos establecer de manera cuantitativa las diferentes contribuciones al fondo. Esto facilita la identificación de los fondos a reducir, mediante el establecimiento de cortes cinemáticos. Además estas simulaciones nos permiten guardar las cadenas de desintegración y con ellas las variables de generación.

Por el comportamiento de nuestro proceso, consideramos sólo eventos que cumplen una serie de condiciones para pertenecer a la región de señal (región rica en señal en la que realizaremos nuestra medida); se ha de exigir la presencia en el estado final de 3 leptones ligeros de los que 2 de ellos formarán un par OSSF, y un p_T^{miss} mayor que 50 GeV ya que esperamos partículas en el estado final que no pueden ser detectadas. Es posible entonces catalogar los eventos en función de la masa invariante del par OSSF, M_{ll}, para separar aquellos procesos que incluyan un Z en la cadena de desintegraciones (Z *'on-shell'*) de los que no.

Vamos a considerar tres puntos de señal, es decir, tres valores para $m_{\mathrm{NLSP}} - m_{\mathrm{LSP}}$ (NLSP hace referencia a *Next to Lightest Supersymmetric Particle*, es decir $\tilde{\chi}_1^{\pm}$ o $\tilde{\chi}_2^0$) como cota superior energética de los productos de la desintegración de Z o W en nuestro estado final. Tras aplicar los requisitos previos, la distribución de M_{ll} datos, fondo y señal tiene la forma de la figura 9.

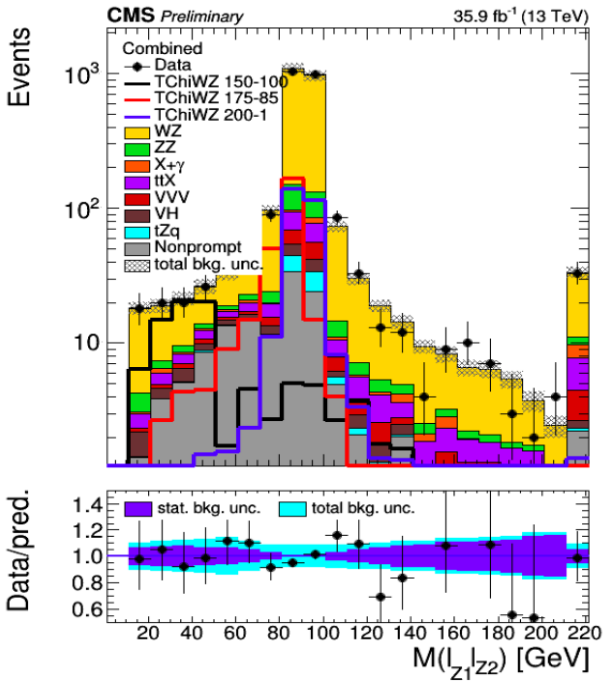

Figura 9: Distribución M_{ll} sin cortes cinemáticos

La figura 9 es un histograma apilado que incluye los eventos de las consideradas posibles señales (representadas por líneas), fondo (en el que se apilan las posibles fuentes para poder comparar visualmente con los datos) y datos (representados por puntos). El gráfico de debajo del histograma apilado nos indica las incertidumbres estadísticas y totales de la simulación de fondo. En esta figura 9 observamos dos regiones de interés para nuestra señal (además del espectro completo, que también es de interés): $M_{ll} < 50\,GeV$ y $M_{ll} \sim m_Z$. El último bin contiene el número de eventos superiores a $M_{ll} = 220$ GeV.

Para la obtención de éste y los siguientes histogramas apilados se usarán cortes cinemáticos que nos proporcionan una calidad mínima de evento y cortes para asegurarnos de que los eventos pasan el trigger, que haya un estado final con 3 leptones ligeros con la suficiente calidad y que haya un par OSSF. Además se establecerán los cortes cinemáticos para asegurarnos de que estudiamos la región de interés. En el caso de la figura 9, como se representa el espectro completo no especificamos otros cortes.

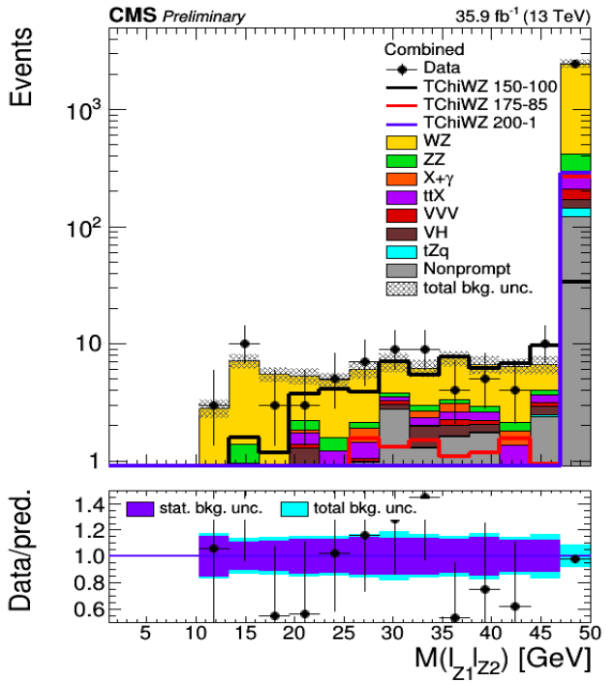

Figura 10: Región de señal $M_{ll} < 50\,GeV$

En la figura 10 (correspondiente a la región de interés $M_{ll} < 50\,GeV$) podemos ver que el punto de señal TChiWZ(150-100) nos da una señal con valores similares al fondo. Sería interesante entonces estudiar esta región en profundidad. Conviene aclarar que el último bin del histograma apilado representado en la figura 10 contiene eventos por encima de $M_{ll} = 50$ GeV, que despreciaremos en el estudio de esta región.

Respecto a la región que incluye al Z en la cadena de desintegraciones ($M_{ll} \sim m_Z$) en la figura 9 vemos que surge un problema, ya que es la misma región en la que se encuentra el grueso del fondo WZ. Queda entonces patente que caracterizar este fondo es de extrema importancia. Para esto hacemos uso de otra variable de interés, la masa transversa del leptón que no pertenece al par OSSF, que llamaremos $M_T = \sqrt{2p_T^{miss}p_T^l(1-\cos\Delta\phi)}$. En procesos que involucran la producción de W, esta M_T caerá de manera pronunciada en valores superiores a la masa del W ($\sim$ 80 GeV), como se observa en la figura 11.

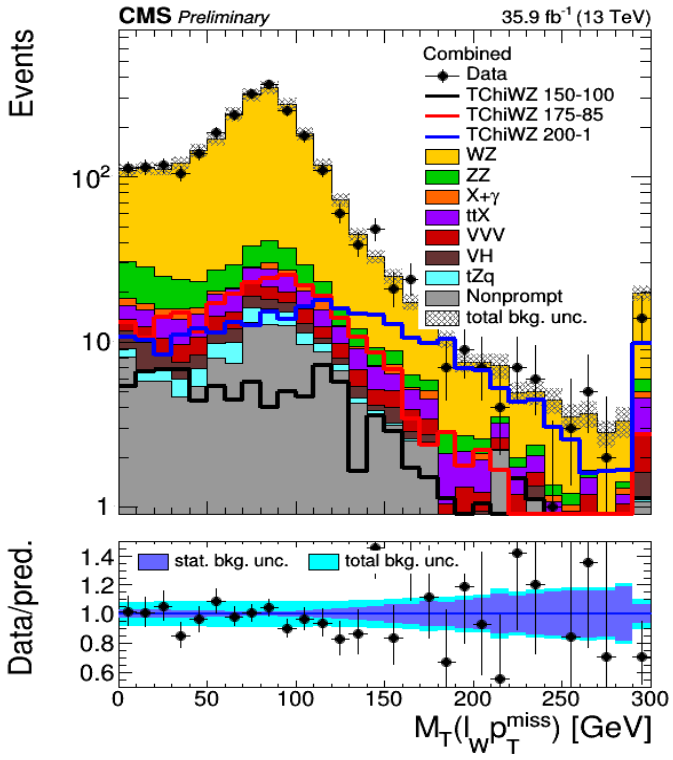

Figura 11: Distribución de M_T con caída en torno a 80 GeV

Experimentalmente hay una cierta ambigüedad al asignarle M_T y M_{ll} a un evento: en el proceso hay 3 leptones que comparten sabor, de los cuales el par OSSF que tiene una masa invariante similar a Z se identifica con M_{ll} y el leptón restante se toma como producto del W y se usa para el cálculo de M_T. Cabe entonces la posibilidad de que asignemos erróneamente al par OSSF el leptón que proviene del W, si este forma una masa invariante más parecida a la m_Z junto con uno de los leptones que provienen del Z/γ^*. Además se ha de tener en cuenta que el fondo WZ es una combinación de procesos WZ y Wγ^*; para sucesos Wγ^* no tiene sentido esperar observar un par OSSF con masa invariante similar al Z.

Si nos fijamos en la cola de la distribución de la masa transversa del leptón no perteneciente al OSSF ($M_T > 160$ GeV), representada en la figura 12:

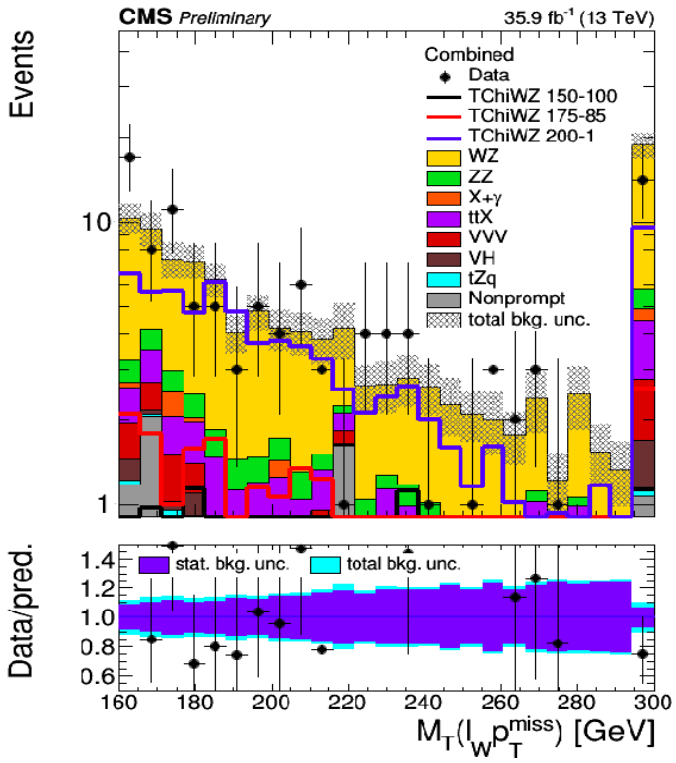

Figura 12: Cola de la distribución M_T

Tenemos predichos $S = 111 \pm 4$ eventos de señal y $B = 112 \pm 3$ eventos de fondo para $D = 100 \pm 10$ datos obtenidos. Esto hace una $S/B = 1{,}00 \pm 0{,}05$. Esta será nuestra principal región de estudio.

La distribución M_T del W decae bruscamente al superar m_W, pero cuando hay etiquetado de leptones erróneo (se asocia M_T a uno de los leptones provenientes de Z o γ^*), la distribución de M_T no tendrá esa característica caída de manera completa y se podrán obtener sucesos con M_T muy alto. Otra causa de eventos con alto M_T es una mala medida de p_T^{miss} y su dirección, pero el efecto es mucho menor y se considera secundario [6].
En relación al etiquetado erróneo de leptones, podemos asignar leptones de generación a los leptones reconstruidos buscando en los leptones a nivel de generación el más cercano en términos de ΔR . De esta manera podemos hacer una división entre eventos bien o mal etiquetados.

Si aplicamos un cambio de etiquetado en los eventos en los que se considera erróneo, se puede observar una distribución del M_{ll} muy parecida a la que cabría esperar para desintegraciones de γ^*. Esto parece indicar que el algoritmo que decide el etiquetado de los leptones fuerza al M_{ll} a valores compatibles con m_Z en un eventos Wγ^*, ocasionando un mal etiquetado y haciendo que el M_T de estos eventos no esté limitado por m_W.
Aparece una resonancia en $M_{ll} = m_Z$ para los eventos etiquetados por el algoritmo; esto es debido a que

se toman como mal etiquetados eventos que realmente son WZ y estaban bien etiquetados. Cabe entonces el desarrollo de un nuevo algoritmo que facilite el etiquetado con un menor porcentaje de fallos, para una reducción del fondo debido al mal etiquetado y por tanto una mejora en la intensidad de señal observada.

5. Método de estudio del reetiquetado de leptones

5.1. Aprendizaje automático

El aprendizaje automático (*machine learning*) es una rama de la inteligencia artificial que consiste en el estudio de algoritmos que sean capaces de mejorar su propio desempeño mediante experiencia, considerando esta como la exposición a información contenida en datos.

Podemos hacer una distinción en tres grandes clases:

- **Aprendizaje supervisado**: consiste en algoritmos que aprenden de un conjunto de datos con clases conocidas, e intentan generalizar al conjunto de todos los datos de entrada posibles.
- **Aprendizaje no supervisado**: las clases en las que se pueden dividir los datos no están previamente conocidas.
- **Aprendizaje mediante refuerzo**: los algoritmos aprenden mediante críticas provenientes de un agente. Estas críticas proveen información sobre la calidad de la solución, pero no dan información sobre como mejorar el algoritmo. La mejora se obtiene explorando iterativamente el espacio de soluciones.

Nos enfrentamos con un problema de clasificación binaria, que afrontamos valiéndonos de aprendizaje supervisado. Partiendo de esto, hay dos formas de abordar el problema de clasificación: considerar que tenemos suficiente conocimiento del conjunto de datos como para afirmar que algunas características de este son más relevantes que otras o no hacer esta distinción, asumir nulo conocimiento de las relaciones entre características y tomar el mayor número de características posible. En nuestro caso haremos una preselección para asegurar que los datos con los que trabajamos son la señal de los procesos estudiados.

En resumidas cuentas, esperamos diseñar un algoritmo que valiéndose de un volumen elevado de datos automatice la búsqueda de relaciones entre características del conjunto de eventos y que nos permita realizar una clasificación dentro de unos márgenes de error aceptables. Como veremos a continuación, este es el escenario perfecto para aplicar *deep learning*.

5.2. *Deep learning*

El aprendizaje profundo (*deep learning*) es una rama del *machine learning* que se basa en redes neuronales artificiales (*Artificial Neural Networks*, ANN) y que nos permite afrontar problemas complejos mediante una estructura profunda de capas formadas por neuronas.

Las redes neuronales artificiales intentan simular la forma de aprendizaje de organismos biológicos. De la misma forma que en el sistema nervioso de los organismos biológicos, las redes neuronales contienen unidades de computación que denominamos nodos o neuronas por su similitud en estructura.

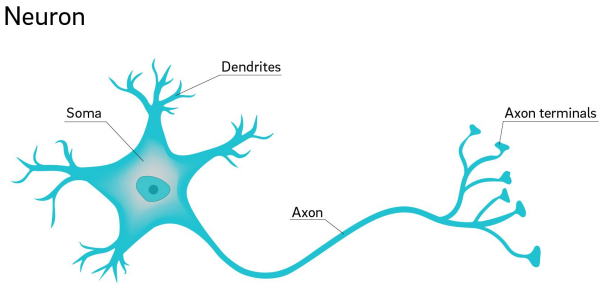

Figura 13: Neurona natural

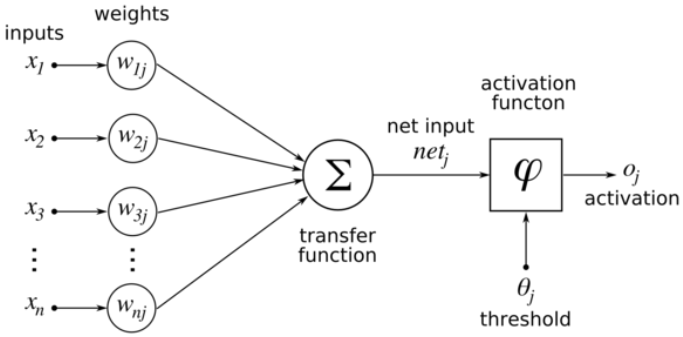

Figura 14: Estructura de neurona artifical

Estas neuronas interaccionan entre ellas mediante pesos, que vienen a representar la fuerza de la conexión establecida entre ellas; la entrada que recibe una neurona (la salida de otra neurona previa o los datos de entrada) es escalada por un peso y afecta a la función computada en esa neurona.

A grandes rasgos, el aprendizaje que conseguimos ocurre gracias a cambios en los pesos que conectan neuronas. De esta manera, al modificar los pesos la función computada será modificada también, permitiendo hacer predicciones más correctas.

5.3. Redes neruronales artificiales: estructura y proceso de aprendizaje

5.3.1. Perceptrón

La red neuronal más simple, conocida como perceptrón y representada en la figura 15, contiene una sola capa de entrada y un nodo de salida. La conexión desde la entrada a la salida de cada variable asociada a característica está modulada por un peso por el que se multiplican y suman en el nodo de salida. Posteriormente se aplica una función de activación para realizar una predicción de clase.

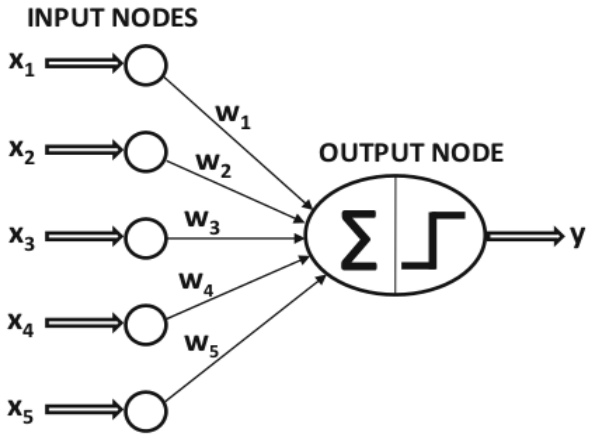

Figura 15: Estructura de un perceptrón [8]

Pongamos un ejemplo: consideremos una instancia de aprendizaje $(\mathbf{X}, y)$, donde $\mathbf{X} = [x_1, ...x_d]$ contiene variables asociadas a características de evento e $y \in \{-1, +1\}$ es el valor observado de la clase, esto es, el que nos viene dado como parte de los datos de entrenamiento y que debemos predecir para casos sin clasificar.
La capa de entrada (en la que no se produce ningún tipo de computación) contiene d nodos, correspondientes a las d características $\mathbf{X}$ a las que son asociadas d conexiones con peso $\mathbf{W}$ hacia un nodo de salida. En este nodo de salida se computa la función $\mathbf{W} \cdot \mathbf{X} = \sum_{i=1}^{d} w_i \cdot x_i$, que nos proporciona un valor real del cual tomaremos la función signo como activación para predecir la variable dependiente asociada al conjunto de características $\mathbf{X}$, a la que llamaremos $\hat{y}$. Cabe añadir que usamos la función signo por simplicidad; se pueden usar otras funciones de activación (tangente hiperbólica, función sigmoide...) dependiendo de los objetivos del estudio.
El error o función de pérdida en la clasificación será entonces $E(\mathbf{X}) = y - \hat{y}$. Como $y, \hat{y} \in \{-1, +1\}$ el error está comprendido en $\{-2, 0, +2\}$. En los casos en los que el error no sea nulo, los pesos de las conexiones han de ser modificados en dirección opuesta al gradiente del error. Específicamente, el vector $\mathbf{W}$ se modifica de acuerdo a la relación $\mathbf{W} \Leftarrow \mathbf{W} + \alpha(y - \hat{y})\mathbf{X}$ donde α es un parámetro que regula el ratio de aprendizaje de la red.

El perceptrón itera el procedimiento para todas las muestras una a una o por grupos (*batches*[6]), pudiendo usar una muestra de entrenamiento más de una vez. Cada una de estas iteraciones se conoce como *época*, y en principio se repite el proceso indefinidamente hasta hallar convergencia[7], tener un error menor que un límite establecido o llevar a cabo un número de iteraciones determinado.

Si los datos no son linealmente separables, no se garantiza convergencia hacia error nulo, por lo que se ha de recurrir a otros métodos o a estructuras neuronales más complejas.

5.3.2. Perceptrón multicapa (MLP)

Estructura del MLP

Como su nombre indica, el perceptrón multicapa posee una estructura que consta de más de una capa en la que ocurre computación: están formados por una capa de entrada, una serie de capas intermedias (*hidden layers*) y una capa de salida. Las capas intermedias se denominan como 'ocultas' porque las computaciones que tienen lugar en ellas no son accesibles al usuario. Este tipo de estructura se conoce como red *feed-forward* ya que todos los nodos de una capa está conectados con los de la capa siguiente y la información sobre la predicción se transmite en dirección entrada-salida, como se observa en la figura 16.

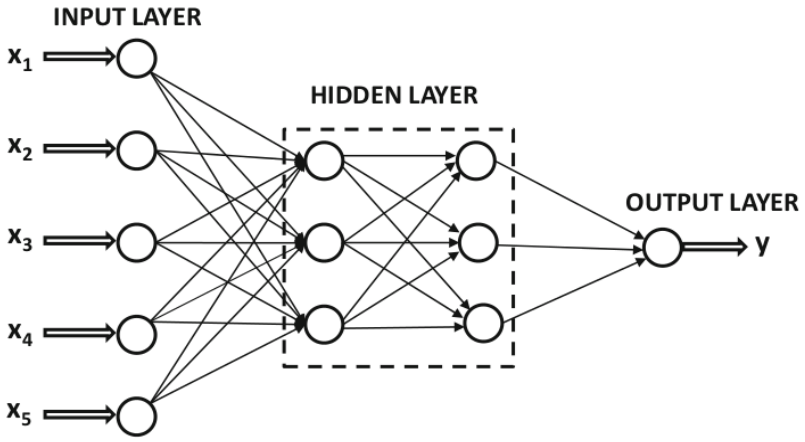

Figura 16: Estructura de un perceptrón multicapa con capa de entrada, dos capas ocultas y una capa de salida [8]

[6] $\mathbf{W} \Leftarrow \mathbf{W} + \sum_{\mathbf{X} \in S} \alpha(y - \hat{y})\mathbf{X}$ donde α es el ratio de aprendizaje y S es un subconjunto de instancias de aprendizaje
[7] Converge a error nulo cuando los datos son linealmente separables [11]

Si la red neuronal contiene n_k nodos en cada una de sus k capas, entonces el vector de salida de cada capa será $\mathbf{h}_k$ con dimensión n_k, esto es, nos referiremos a partir de ahora al número de nodos en cada capa como la dimensión de esa capa. Las conexiones entre capas están moduladas por pesos, y podemos definir el conjunto de pesos de las conexiones entre dos capas adyacentes, digamos las capas i e $i+1$ por una matriz P_i de dimensión $n_i \times n_{i+1}$.

El flujo de información en la época T es el siguiente: el vector datos de entrada $\mathbf{x}$ pasa a la primera capa oculta como $\mathbf{h}_1 = \Phi(P_1^T \mathbf{x})$, el vector de salida de la capa oculta $\mathbf{h}_i$ será el de entrada de la capa oculta $\mathbf{h}_{i+1}$, es decir $\mathbf{h}_{i+1} = \Phi(P_{i+1}^T \mathbf{h}_i)$ donde $i \in \{1...k-1\}$, la entrada de la capa de salida será $\mathbf{o} = \Phi(P_{k+1}^T \mathbf{h}_k)$ y finalmente la capa de salida nos proporcionará $\hat{y}$.
Se hace referencia a las funciones de activación de manera genérica, ya que algunas son aplicadas elemento a elemento y otras directamente sobre los vectores, y asumimos implícitamente que todos los nodos de una misma capa usan la misma función de activación. Esto es verídico para la gran mayoría de estructuras neuronales (como es nuestro caso), en las que además la misma función de activación se comparte incluso para todas las capas ocultas.
Otra consideración que debemos hacer es si la estructura está totalmente conectada; más adelante veremos cómo puede ayudarnos establecer un *dropout* entre capas, es decir, nulificar de forma aleatoria algunas entradas de las matrices peso, P_i, para evitar fenómenos no deseados en el aprendizaje de la red.

Entrenamiento del MLP

El entrenamiento de este tipo de red es más complejo que el del perceptrón simple, ya que la función de pérdida es una función compuesta de los pesos de las capas. El gradiente de esta se computa mediante el denominado algoritmo de retropropagación, que consta de dos fases:

- **Fase hacia delante**: se introduce la instancia de aprendizaje en la red y se propaga la información hacia delante, debidamente pesada. La predicción final se compara con la información de clase de la instancia de aprendizaje y se calcula la derivada de la función de pérdida respecto a la predicción.

- **Fase hacia atrás**: El objetivo es hallar el gradiente de la función de pérdida respecto a los pesos, y esto se hace mediante la regla de la cadena multivariable. Esta fase se denomina 'hacia atrás' porque es la forma en la que se van hallando los gradientes, que serán usados posteriormente para modificar los pesos y reducir el valor de la función de pérdida.

Esta segunda fase de la retropropagación es algo compleja con lo que vamos a introducirla con un caso sencillo para ilustrar la aplicación sucesiva de la regla de la cadena que se usa en esta segunda fase de la retropropagación: tenemos un nodo de entrada, dos nodos en una capa oculta y un nodo de salida. Considerando conexos todos los nodos de capas adyacentes, tenemos dos recorridos posibles entre el nodo de entrada y el de salida.

1. El nodo de entrada computa una función $f(p)$ respecto al peso de entrada p.
2. El nodo que corresponde al primer recorrido (asignado arbitrariamente) tiene como entrada $x = f(p)$ y computa una función $g(x)$. El otro nodo que corresponde al otro recorrido tiene como entrada $y = f(p)$ y computa una función $h(y)$.
3. El nodo de salida tiene como entradas $n = g(x)$ y $m = h(y)$, y computa una función $j(n, m)$
4. La salida será $s = j(n, m) = j(g(f(p)), h(f(p)))$

Entonces, veamos cómo varía la salida respecto al peso de entrada. Aplicando la regla de la cadena multivariable:

$$\frac{\partial s}{\partial p} = \frac{\partial s}{\partial n}\frac{\partial n}{\partial p} + \frac{\partial s}{\partial m}\frac{\partial m}{\partial p} \tag{1}$$

Ahora aplicando la regla de la cadena:

$$\frac{\partial s}{\partial p} = \frac{\partial s}{\partial n}\frac{\partial n}{\partial x}\frac{\partial x}{\partial p} + \frac{\partial s}{\partial m}\frac{\partial m}{\partial y}\frac{\partial y}{\partial p} = \frac{\partial j(n,m)}{\partial n}\, g'(x)\, f'(p) + \frac{\partial j(n,m)}{\partial m}\, h'(y)\, f'(p) \tag{2}$$

El primer sumando se corresponde con el primer recorrido, y el segundo con el otro recorrido posible. Lo que se observa es que la derivada de la salida respecto al peso de entrada es la suma en todos los caminos de los productos de las derivadas parciales de las salidas de todos los nodos pertenecientes a un camino dado.

Ahora pasemos a considerar el caso general, mucho más complejo, en el que:

- Tenemos i nodos ocultos o seguidos y una salida s.
- El peso de la conexión entre o_r a o_{r+1} es $p_{(o_r, o_{r+1})}$
- Existe un conjunto de recorridos C entre un nodo o_r y s.
- Tenemos una función de pérdida P que se computa respecto a s.

Entonces se obtene el gradiente de la función de pérdida respecto a los pesos como:

$$\frac{\partial P}{\partial p_{(o_{r-1},o_r)}} = \frac{\partial P}{\partial o_r}\frac{\partial o_r}{\partial p_{(o_{r-1},o_r)}} = \frac{\partial P}{\partial s}\left[\sum_{[o_r,o_{r+1},\ldots,o_i,s]\in C}\frac{\partial s}{\partial o_i}\prod_{j=r}^{i-1}\frac{\partial o_{j+1}}{\partial o_j}\right]\frac{\partial o_r}{\partial p_{(o_{r-1},o_r)}} \tag{3}$$

El término $\frac{\partial o_r}{\partial p_{(o_{r-1},o_r)}}$ se obtiene de la siguiente manera:

$$\frac{\partial o_r}{\partial p_{(o_{r-1},o_r)}} = o_{r-1}\cdot\Psi'(\sigma_{o_r}) \tag{4}$$

siendo σ_{o_r} una combinación lineal de las entradas al nodo oculto o_r y $\Psi'(\sigma_{o_r})$ haciendo referencia a la variación del resultado de la función de activación de o_r respecto a estas entradas.
El término $\frac{\partial P}{\partial o_r}$ es relativamente más complejo, ya que como se puede observar aglutina cada posible recorrido y crece de forma exponencial respecto a la longitud de estos. Para su obtención se usan técnicas de programación dinámica que simplifican el cálculo de esta ingente cantidad de operaciones.

Del mismo modo que en el perceptrón más simple, este proceso de aprendizaje se repite en épocas hasta llegar a la convergencia[8] o detener el proceso en un punto determinado (número de épocas determinado, valor de la función de pérdida determinado...).

[8]El teorema de aproximación universal dice que una red *feed-forward* con una sola capa oculta que contenga un número finito de neuronas y con ciertas condiciones en su función de activación puede aproximar cualquier función continua de un subconjunto compacto de $\mathbb{R}^n$ [9]

6. Aplicación y resultados

Los apartados de esta sección son altamente interdependientes (al implementar el método surgen unos ciertos problemas y necesidades, que se han de solucionar modificando ciertos hiperparámetros[9], que puede hacer surgir otros problemas...), por ello se ha estructurado de la siguiente forma: primero se trata la implementación de forma general, después los problemas que han surgido durante el entrenamiento (y sus posibles soluciones) y finalmente se profundizará en el modelo final y los resultados conseguidos.

El entrenamiento de la red nuronal se ha llevado a cabo en Python, usando de librerías **ROOT** para la lectura de los datos, **math** y **numpy** para cálculos y creación/manipulación de arrays, **matplotlib** para gráficas y **Tensorflow** (usando la interfaz **keras**) para incorporar la red neuronal y su aprendizaje.

6.1. Adaptación del método

El objetivo es reducir lo máximo posible el fallo de etiquetado de los leptones. Para esto se va a realizar una reclasificación de los eventos respecto a su etiquetado de leptones, actualizando los resultados de la clasificación previa por variables de generación para generar clases bien/mal etiquetadas.

En los referente al tratamiento previo de datos, se van a separar en dos grupos: uno de ellos (cerca de un tercio de la cantidad total de eventos) servirá para entrenar la red y el otro grupo será usado para comprobar que el entrenamiento no es demasiado específico (uno de los principales problemas en el entrenamiento, denominado sobreajuste, en el que se profundiza en el apartado siguiente). Además, la muestras Monte Carlo que usaremos para entrenar la red y posteriormente evaluar los resultados corresponden a distintos generadores (aMC@NLO y POWHEG repectivamente), por lo que eliminamos posibles sesgos al ser totalmente independientes.
También será necesario establecer los cortes cinemáticos que nos permitan asegurar que los eventos que tratamos son los de interés, y usaremos como entradas de la red todas las características de evento relevantes, sin hacer distinciones sobre la importancia relativa de cada una.

Dado que en la gran mayoría de eventos con los que vamos a entrenar la red el etiquetado de los leptones era correcto ($\sim 93\,\%$), estamos tratando con un conjunto de datos con **clases desequilibradas**; se puede incorporar entonces un peso a cada clase con el objetivo de normalizarlas (en lo respectivo al entrenamiento y representaciones gráficas posteriores) y que esta característica del conjunto de datos no nos presente mayores problemas.

[9] Los hiperparámetros son valores fácilmente accesibles y ajustables que sirven para controlar la forma de aprendizaje de la red neuronal.

Para la estructura de red se ha usado el modelo secuencial de keras[10], que permite, como su nombre indica, la creación de una pila sencilla de capas, siendo la salida de una la entrada de la siguiente. Esta API nos permite la fácil modificación de hiperparámetros como el tipo de capa, el número de entradas de cada capa, la función de activación...
De la misma forma, el entrenamiento es altamente personalizable respecto a épocas, función de pérdida usada y demás características.

6.2. Fenómenos en el entrenamiento de la red

Sobreajuste

Lo que prentendemos al entrenar una red, como se ha comentado en apartados anteriores, es alcanzar la capacidad de generalizar lo aprendido en el conjunto de datos usado sobre el total de datos posibles.
El sobreajuste, más comúnmente conocido como *overfitting*, aparece cuando la red se entrena demasiado sobre un conjunto de datos. Esto causa que la red considere rasgos específicos de los datos usados como determinantes en el resultado de la clasificación cuando pueden no serlo de forma general (ocurre con más frecuencia en conjuntos de datos pequeños respecto a las características de evento).

Dado que el sobreajuste es un problema intrínseco al entrenamiento, existen ciertas técnicas para identificarlo o evitar causarlo en la medida de lo posible. Explícitamente se ha tenido que hacer uso de:

- ***Test/train split***: La parte más peligrosa del sobreajuste es que, al evaluar el desempeño de la red en los datos sobre los que se ha entrenado y dar excelentes resultados, nos puede hacer creer que su capacidad clasificatoria sobre otro conjunto cualquiera de datos del proceso estudiado será también muy buena. Para evitar confusiones de este tipo se realiza el llamado *test/train split*: dividimos el conjunto de datos que tenemos en un subconjunto sobre el que se entrena la red (*train set*) y otro que se usa exclusivamente para evaluar el correcto funcionamiento de la red (*test set*). De esta manera podemos comparar la evaluación de la clasificación y la evolución de los valores para la funcion de pérdida en ambos subconjuntos para comprobar que no hay sobreajuste.

  En nuestro caso la implementación de esta división es factible, ya que contamos con una gran cantidad de eventos y reducir los datos sobre los que se entrena la red no afecta en gran medida.

- **Incremento de profundidad**: Redes con más capas tienen funciones compuestas de pesos más largas. Esto se puede considerar una forma de regularización[11] ya que, debido al productorio de la regla de la

[10] Documentación de keras en https://keras.io/api/
[11] Proceso que reduce el valor de determinados coeficientes o los hace tender a cero.

cadena en la retropropagación, las capas más superficiales imponen la importancia relativa de su variación sobre las más profundas, haciendo más costoso el aprendizaje de la red en algunos casos. Esto puede llegar a ser un inconveniente si llega a casos extremos, conocidos como desvanecimiento de gradiente o, en caso contrario, gradiente explosivo: los cambios en los pesos son demasiado pequeños o grandes (respectivamente) debido a que las derivadas en las capas más externas son demasiado pequeñas o grandes. Estos problemas son muy dependientes de la función de activación escogida.

- ***Dropout***: Es una técnica de regularización que consiste en anular un porcentaje determinado de pesos entre capas determinadas de forma aleatoria. Esto puede considerarse una forma de "promediar", pero conlleva una disminución en la reproducibilidad del modelo. En cuanto a la implementación de esta técnica se ha tenido cuidado, ya que *dropouts* excesivos hacen que la red necesite más épocas para llegar a valores similares de la función de pérdida, alargando el tiempo de computación.

- ***Early stopping***: Método de regularización que compara el valor de la función de pérdida de una época con épocas anteriores, y si este valor ha ido en aumento, hace que se detenga el entrenamiento de la red. Es útil ya que nos ayuda a reducir el sobreajuste que puede darse en épocas tardías.

Lentitud en la convergencia

Un problema de las redes profundas es el tiempo de computación del entrenamiento. De manera similar al desvanecimiento de gradiente, hay situaciones en las que se sufren las consecuencias de que en la retropropagación los gradientes de unas capas sean computados junto con las demás capas y los pesos no se modifiquen lo suficientemente rápido cuando no hay peligro de sobreajuste. Es por esto que no podemos hacer la red arbitrariamente profunda y debemos de considerar cuál es la mejor relación de resultados contra tiempo de computación.

6.3. Elección de hiperparámetros para el modelo final

El criterio de calidad que establecemos será el mínimo error de etiquetado (*mistag rate*) que puede conseguir la red para un cierto valor de corte en la puntuación entre 0 y 1 (los eventos cuya puntuación sea inferior a este corte serán considerados como "mal etiquetados"[12], y se actualizará su clase para reflejar este hecho) asignada al evento, siendo el valor original de fallo en etiquetado de **2,79 %**.

[12]Cabe señalar que el error en la clasificación incluye tanto los eventos mal etiquetados considerados como bien etiquetados, como los bien etiquetados considerados como erróneos.

Elección de variables de entrada

La elección de las variables de entrada de la red es uno de los factores más determinantes en la calidad del modelo final. Es por esto que escogemos las variables más significativas, dentro de las cuales hay tanto variables asociadas a leptón como variables globales.

Las variables asociadas a cada uno de los leptones del estado final usadas serán:

- **pt**: Momento transverso del leptón.
- **phi**: Angulo azimutal del leptón.
- **eta**: Pseudorapidez del leptón, $\eta = -log(tan(\theta/2))$
- **mass**: Masa del leptón.
- **conePt**: Momento transverso corregido por el aislamiento relativo del leptón.
- **dxy**: Distancia en el plano transverso de la trayectoria del leptón extrapolada hasta el punto más cercano a la colisión.
- **dz**: Distancia en el plano longitudinal de la trayectoria del leptón extrapolada hasta el punto más cercano a la colisión.
- **mva**: Discriminante multivariable que se usa para decidir si el leptón cumple la selección *prompt*.
- **jetDR**: Distancia angular del leptón al jet más cercano.
- **ptratio**: Ratio entre el momento transverso del leptón y el momento transverso del jet más cercano.
- **sip3d**: Distancia 3D entre el punto de colisión y el punto más cercano al mismo en la trayectoria del leptón dividido por la incertidumbre de esta cantidad.
- **miniRelIso**: Aislamiento relativo del leptón.

Las variables correspondientes a la cinemática de todo el evento serán:

- **mll_3l**: Es lo que nosotros denominamos M_{ll}.
- **mT_3l**: Es lo que nosotros denominamos M_T.
- **deltaR_WZ**: Distancia angular entre el W y Z reconstruidos.
- **wzBalance_pt**: Diferencia entre el momento transverso reconstruido del Z y el W.
- **wzBalance_conePt**:Diferencia entre el momento transverso corregido por el aislamiento relativo reconstruido del Z y el W.
- **m3Lmet**: Masa invariante del sistema de tres leptones más la energía transversa faltante.

Usaremos 6 variables globales y 12 variables por cada uno de los leptones, que hacen 42 variables totales.
A continuación se representan las distribuciones de algunas de estas variables para las simulaciones con las que entrenamos la red y para las simulaciones y datos con los que evaluaremos el reetiquetado, de esta manera podemos cerciorarnos de que las simulaciones y datos son compatibles.

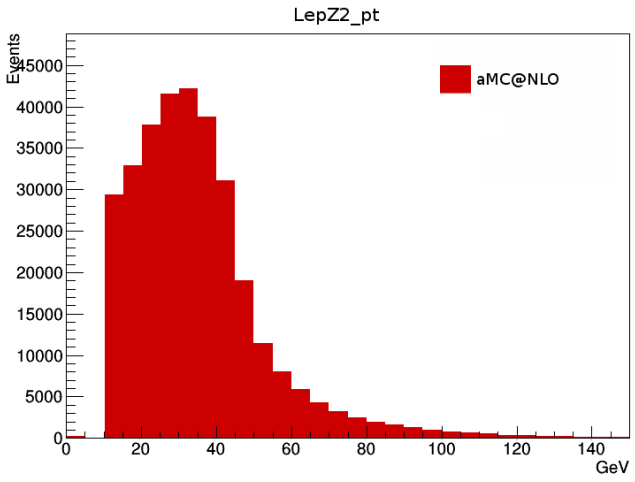

Figura 17: Distribución del momento transverso del leptón Z2 en la simulación de entrenamiento

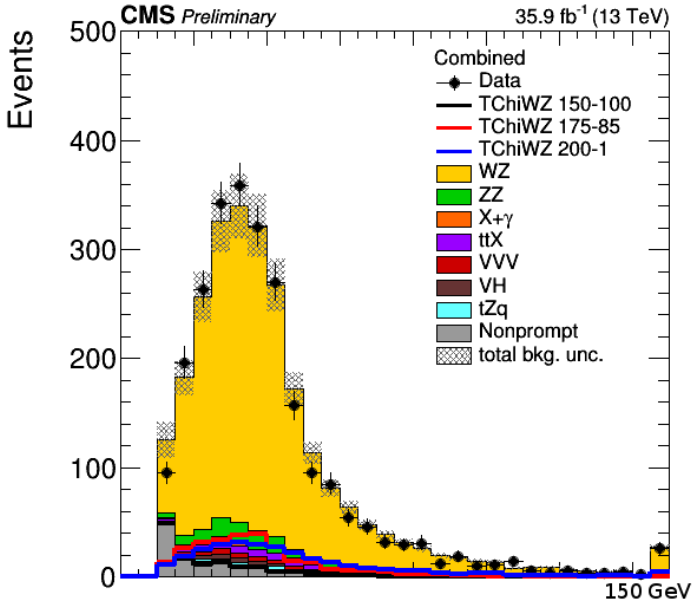

Figura 18: Distribución del momento transverso del leptón Z2 en datos y simulación de análisis

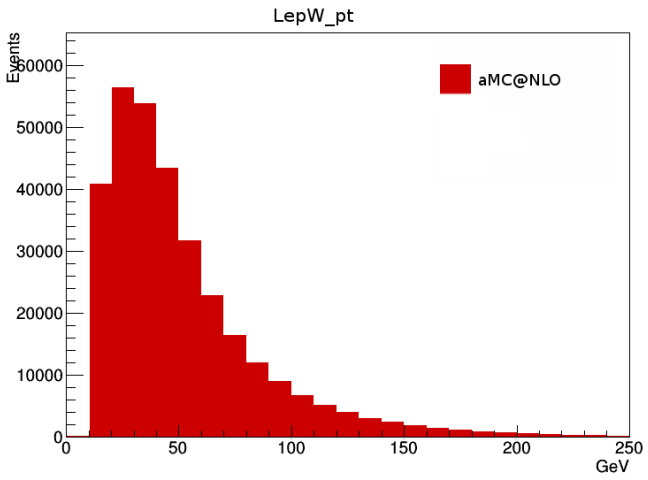

Figura 19: Distribución del momento transverso del leptón W en la simulación de entrenamiento

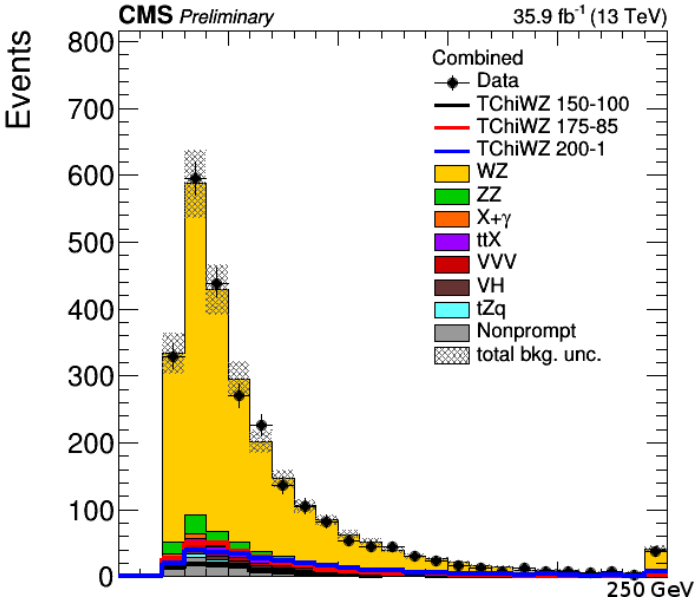

Figura 20: Distribución del momento transverso del leptón W en datos y simulación de análisis

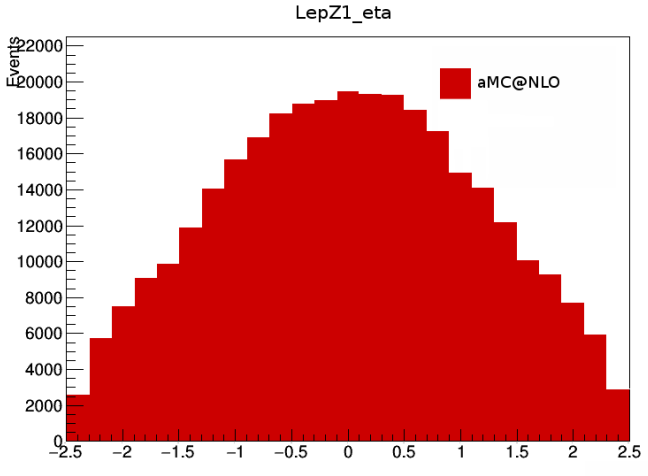

Figura 21: Distribución de la pseudorapidez del leptón Z1 en la simulación de entrenamiento

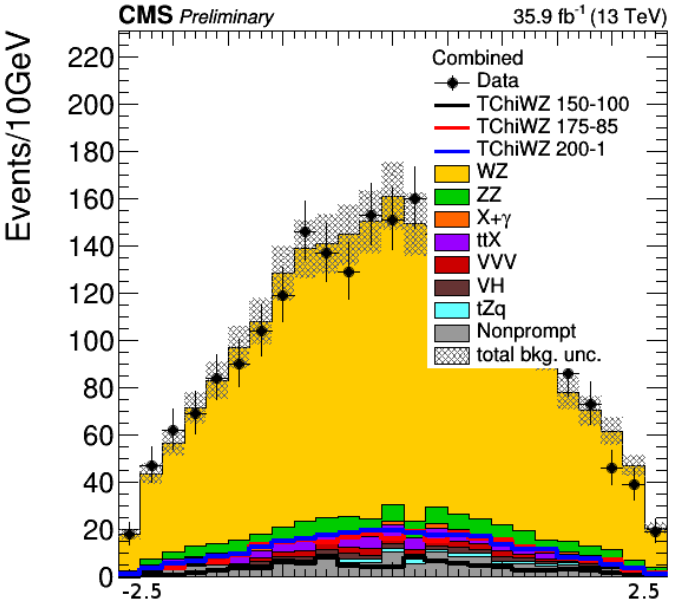

Figura 22: Distribución de la pseudorapidez del leptón Z1 en datos y simulación de análisis

Elección de la funciones de pérdida y activación

La función de pérdida por excelencia para clasificaciones binarias (como es nuestro caso, específicamente con las clases para evento 'leptones mal etiquetados' y 'leptones bien etiquetados') es la **entropía cruzada binaria** [9], que tiene la forma:

$$ECB = -(1/N)\sum_{i=1}^{N} y_i log(p(y_i)) + (1 - p(y_i)) \tag{5}$$

donde $y \in [0, 1]$ es la clase asignada al evento y $p(y)$ es la probabilidad predicha de un evento de ser de esa clase.

La elección de funciones de activación es también sencilla: vamos a usar **ReLU** (*Rectified Linear Unit*) para la capa de entrada y las capas ocultas, y una **sigmoide** como función de activación de la capa de salida.

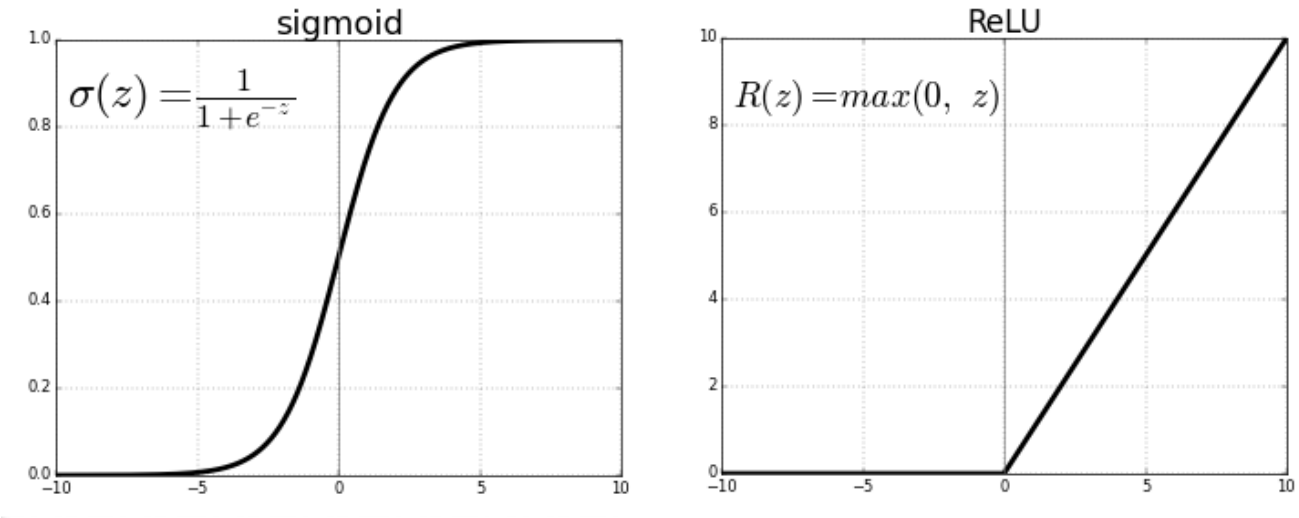

Figura 23: Comparativa en entorno cercano a 0 de sigmoide y ReLU

En la figura 23 se observa que el rango en el que la derivada es cercana a cero es mayor en la sigmoide, esto es, puede llevar a problemas de desvanecimiento de gradiente para un rango mayor de valores que ReLU (esta solo satura para valores menores que 0, y su derivada es 1 para todos los demás). Además, el requerimiento computacional de la sigmoide es mayor, ya que necesita de la computación de un exponente, mientras que ReLU es una simple elección. El único inconveniente que presenta ReLU es que no vamos a poder usarla en la capa de salida, ya que finalmente se necesita un resultado entre 0 y 1 para que sea posible introducirlo en la función de pérdida que hemos escogido.

Elección de optimizador

El **optimizador** es el algoritmo que permite el cambio de pesos una vez que la función de pérdida establece 'cuánto' deben cambiar estos. El **ratio de aprendizaje** es una característica modulable del optimizador, que nos permite decidir cómo de rápido pueden variar los pesos; es una variable delicada así que usaremos el valor por defecto para cada optimizador.

Optimizador	Adagrad	Adadelta	Adam	Adamax
Min. mistag rate	2.74 %	2.63 %	2.59 %	2.58 %

Cuadro 1: Comparativa de error de etiquetado mínimo para optimizadores

Se han comparado 4 optimizadores típicos[13] para clasificador en el cuadro 1 manteniendo el resto de hiperparámetros, para 15 épocas. En esta primera prueba vemos que los mejores resultados de optimizador se obtienen con Adam y Adamax. Esta elección es discutible, ya que el número de épocas es reducido. Tomamos los dos mejores e incrementamos el número de épocas significativamente (~ 70) de tal manera que se asemeje a la clasificación final.

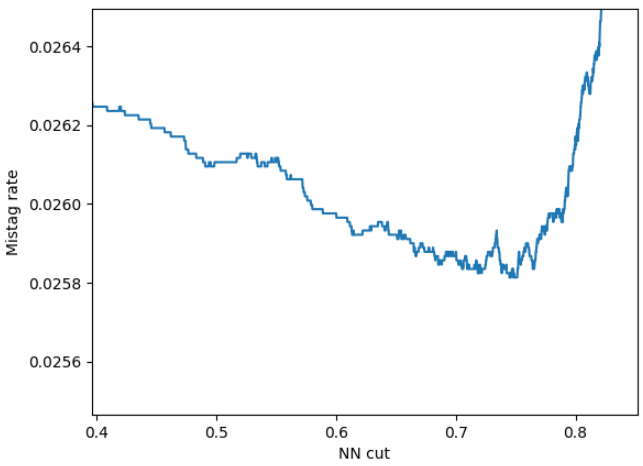

Figura 24: Fallo en etiquetado para Adam

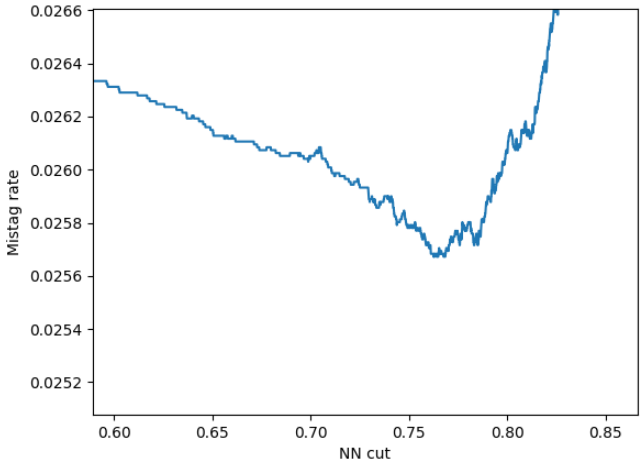

Figura 25: Fallo en etiquetado para Adamax

El eje X de las figuras 24 y 25 representa el corte en la puntuación entre 0 y 1 que le asigna la red a un evento para establecer lo mal/bien etiquetado que está éste, mientras que el eje Y representa el *mistag rate* que tendremos para éste corte. Nos interesa tener el mínimo *mistag rate*, entonces podemos decantarnos por usar Adamax como optimizador para el reetiquetado.

[13]Toda la información sobre estos optimizadores está disponible en *Keras API Reference*, https://keras.io/api/

Elección del número de capas, *dropout*, *batch size* y épocas

Cuando el número de capas es elevado, es conveniente usar *dropouts* como técnica de regularización para evitar el sobreajuste [8]. Se va a usar una red con 4 capas ocultas y *dropouts* entre el 30 % y el 10 % de los nodos, para la capa de entrada y las capas ocultas, además usaremos un número de épocas elevado (75) para obtener el mejor resultado posible. Las elecciones nos vienen facilitadas por el hecho de que el tiempo de computación, en nuestro caso particular, no nos limita.
El *batch size* tomado será de 64; este valor ha sido el que más estabilidad en los resultados nos ha proporcionado.

6.4. Modelo final y resultados de la clasificación

En definitiva, la elección de hiperparámetros para el modelo será la siguiente:

- **Función de pérdida**: Entropía cruzada binaria.
- **Funciones de activación**: Para capas ocultas ReLU, para capa de salida sigmoide.
- **Optimizador**: Adamax.
- **Número de capas y *dropout***: Entrada (30 %), oculta (25 %), oculta (20 %), oculta (15 %), oculta (10 %), salida.
- **Épocas**: 75.
- ***Batch size***: 64.

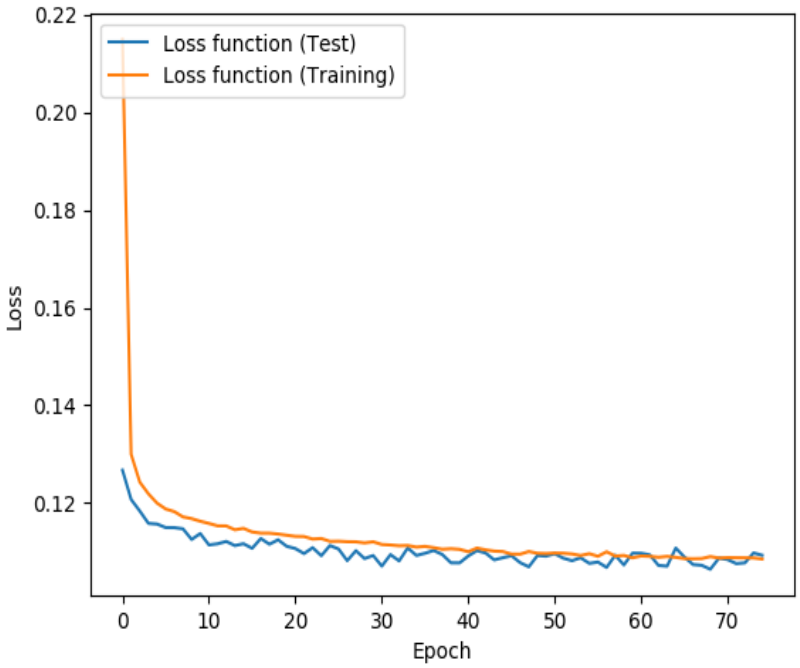

Figura 26: Función de pérdida para *test* y *train*

En la figura 26 representamos el valor de la función de pérdida respecto a las épocas para los grupos de *test* y *train*. Se ve que tienen un comportamiento similar por lo que podemos asegurar que no hay sobreajuste. Obtenemos un *mistag rate* de **2,56 %** para un corte en el valor asignado por la red (recordemos, entre 0 y 1) de **0,81**. El valor original de fallo en etiquetado era de **2,79 %**; hemos conseguido una mejora relativa del **8,2 %**.

Se guardan entonces los pesos del modelo entrenado para aplicar el nuevo etiquetado.

6.5. Aplicación del reetiquetado y comparativa

El paso siguiente será recalcular todas las variables que hemos usado de entrada con el nuevo etiquetado, esperando que se reduzcan de esta manera el fondo debido al mal etiquetado en las regiones de interés.

Estudio de la región sin cortes

En las figuras 27 y 28 se muestran las predicciones para los dos etiquetados con fin comparativo, sin establecer cortes en M_{ll} ni en M_T.

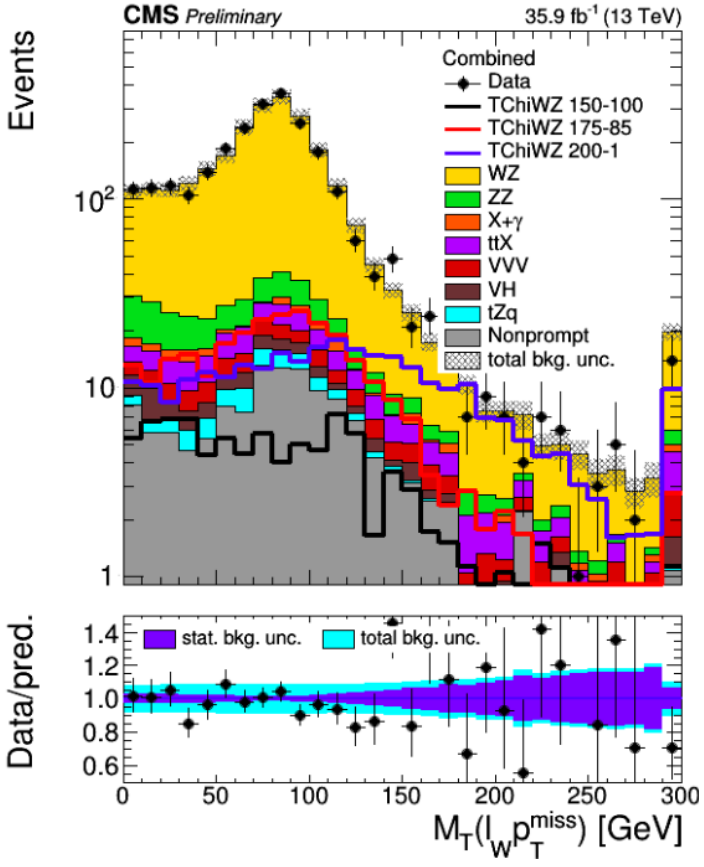

Figura 27: Predicción y datos para el etiquetado original en la región sin cortes

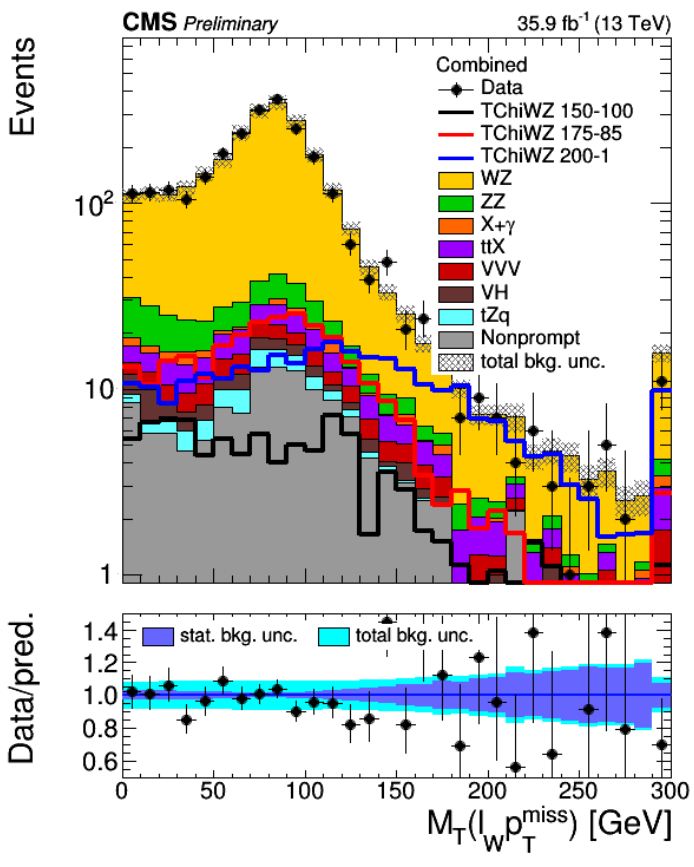

Figura 28: Predicción y datos para el reetiquetado en la región sin cortes

Por las figuras 27 y 28 vemos que datos siguen siendo compatibles con la hipótesis de fondo; además se observa una variación en el número de datos y fondo, esto se debe a que eventos que previamente cumplían con los criterios de selección, una vez cambiadas todas las variables respecto al etiquetado de leptones, dejan de cumplirlos y dejan así de ser considerados.

	N	σ_N
TchiWZ (150-100)	91	6
TchiWZ (175-85)	275	8
TchiWZ (200-1)	289	5
WZ	2100	200
ZZ	120	10
x+γ	20	6
ttX	70	10
VVV	40	20
VH	28	7
tZq	24	8
Nonprompt	130	40

Cuadro 2: Número de eventos y desviación estándar sin cortes establecidos para el etiquetado original

La desviación estándar de N, a la que nos referimos en el cuadro 2 como σ_N viene de:

$$\sigma_N = \sqrt{\text{incertidumbre estadística}^2 + \text{incertidumbre sistemática}^2} \tag{6}$$

donde la incertidumbre estadística es dependiente del número de sucesos observados, se debe a la consideración de fluctuaciones aleatorias en la medida y la incertidumbre sistemática engloba la incertidumbre asociada a los detectores. En nuestro caso particular consideramos las incertidumbres de normalización de la sección eficaz de los procesos. Para el reetiquetado tendremos valores muy similares, como podemos comprobar en el cuadro 3; la reducción de fondo no parece significativa al considerar el espacio de fases completo.

	N	σ_N
TchiWZ (150-100)	92	6
TchiWZ (175-85)	275	8
TchiWZ (200-1)	289	5
WZ	2100	200
ZZ	120	10
x+γ	20	6
ttX	70	10
VVV	40	20
VH	28	8
tZq	24	8
Nonprompt	130	40

Cuadro 3: Número de eventos y desviación estándar sin cortes establecidos para el reetiquetado

A continuación consideramos el punto de señal TChiWZ (150-100) para hallar la relación señal/fondo, que será similar en los otros puntos de señal. Los eventos de fondo predichos son B y los datos obtenidos, D.

	S	σ_S	B	σ_B	D	σ_D	S/B	$\sigma_{S/B}$
Original	92	6	2500	200	2500	50	0.04	0.01
Reetiquetado	92	6	2500	200	2500	50	0.04	0.01

Cuadro 4: Señal, fondo, datos y discriminante señal/fondo con sus incertidumbres para la región sin cortes

Por el cuadro 4 vemos que la relación señal/fondo es prácticamente idéntica para el etiquetado y el reetiquetado sin cortes; hemos de tener en cuenta que no estamos restringidos a la región de interés por la que hemos realizado el reetiquetado.

Definamos la **intensidad de señal observada** como:

$$\mu = \sigma_{obs}/\sigma_{theo} = (D - B)/S \tag{7}$$

con σ_{obs} la sección eficaz que obtenemos y σ_{theo} la sección eficaz esperada. También podemos obtener, para un intervalo de confianza del 95 %, el **límite superior** de la sección eficaz[14], que vendrá dado por:

$$UL(\mu) = \mu + 2\sigma_\mu \tag{8}$$

donde σ_μ hace referencia a la desviación estándar de la intensidad de señal observada.

[14]Resultado hallado en el apartado 3.2 de la referencia [14]

De la misma forma podemos obtener los **límites superiores esperados**: una forma de estimar la incertidumbre si se siguiera directamente el modelo de fondo y no hubiera señal, es decir, si solamente hubiera procesos SM. El cálculo es similar a los límites superiores pero en este caso $D = B$ y $\sigma(D) = \sqrt{B}$.

	μ	σ_μ	Límite superior	Límite superior esperado
TChiWZ (150-100)	-0.4	2	4	5
TChiWZ (175-85)	-0.2	0.8	1.4	1.6
TChiWZ (200-1)	-0.1	0.8	1.4	1.5

Cuadro 5: Intensidad de señal, desviación típica de esta, límite superior y límite superior esperado sin cortes establecidos para el etiquetado original

	μ	σ_μ	Límite superior	Límite superior esperado
TChiWZ (150-100)	-0.4	2	4	5
TChiWZ (175-85)	-0.2	0.8	1.4	1.6
TChiWZ (200-1)	-0.2	0.8	1.4	1.5

Cuadro 6: Intensidad de señal, desviación típica de esta, límite superior (UL) y límite superior esperado sin cortes establecidos para el reetiquetado

Si comparamos los cuadros 5 y 6 no se encuentran diferencias significativas, lo que significa que el proceso de reetiquetado no tiene un efecto significativo en la región de señal total.

Al conocer μ, σ_{theo} para procesos de producción $\tilde{\chi}_1^\pm$ $\tilde{\chi}_2^0$ y sus incertidumbres, podemos predecir la sección eficaz observada para cada punto con su desviación estándar.

	σ_{theo} (fb)	$\sigma_{\sigma_{theo}}$ (fb)	σ_{obs} (fb)	$\sigma_{\sigma_{obs}}$ (fb)	UL(σ_{obs}) (95 % CL)
TChiWZ (150-100)	5180.86	253.22	-2000	10000	20000
TChiWZ (175-85)	2953.28	154.39	-400	2000	4000
TChiWZ (200-1)	1807.39	101.32	-300	1000	3000

Cuadro 7: Secciones eficaces teórica y observada, sus desviaciones estándar y el UL de la sección eficaz observada para la region sin cortes y el etiquetado original

	σ_{theo} (fb)	$\sigma_{\sigma_{theo}}$ (fb)	σ_{obs} (fb)	$\sigma_{\sigma_{obs}}$ (fb)	UL(σ_{obs}) (95 % CL)
TChiWZ (150-100)	5180.86	253.22	-2000	10000	20000
TChiWZ (175-85)	2953.28	154.39	-400	2000	4000
TChiWZ (200-1)	1807.39	101.32	-300	1000	3000

Cuadro 8: Secciones eficaces teórica y observada, sus desviaciones estándar y el UL de la sección eficaz observada para la región sin cortes y el reetiquetado

La μ es negativa, por tanto σ_{obs} será también negativa. Esto, junto a su incertidumbre, indica que tenemos completa compatibilidad con el menor valor que tiene sentido físico, la sección eficaz nula. Se debe a una pequeña reducción en los datos.

Comparando los cuadros 7 y 8 observamos que no hay una variación significativa; el reetiquetado no beneficia a la medida en esta región.

Estudio de la región del espacio de fases restringida por $M_T > 160$ GeV

En las figuras 29 y 30 se muestran las predicciones para los dos etiquetados con fin comparativo, para el corte cinemático $M_T > 160$ GeV.

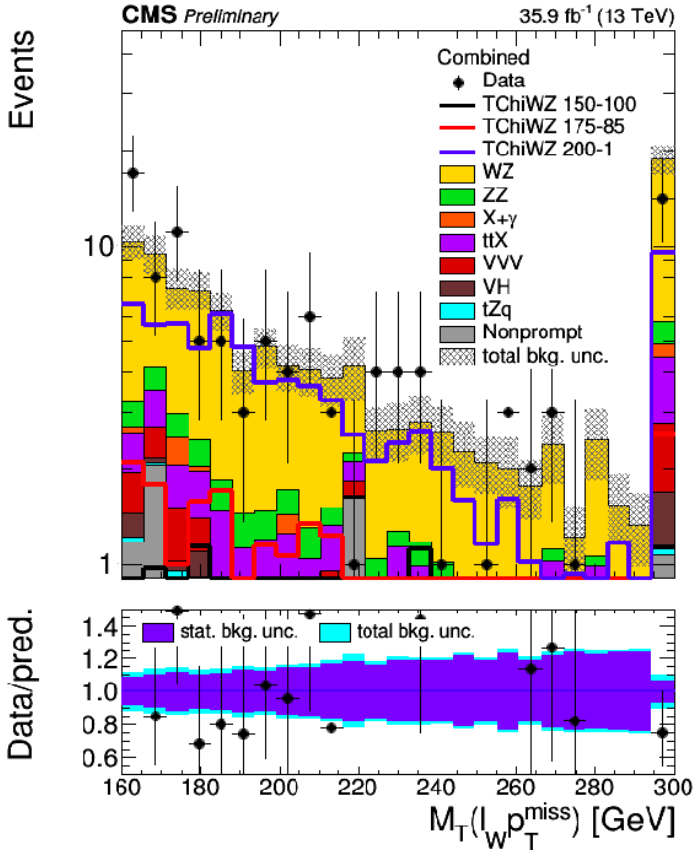

Figura 29: Predicción y datos para el etiquetado original en la región $M_T > 160$ GeV

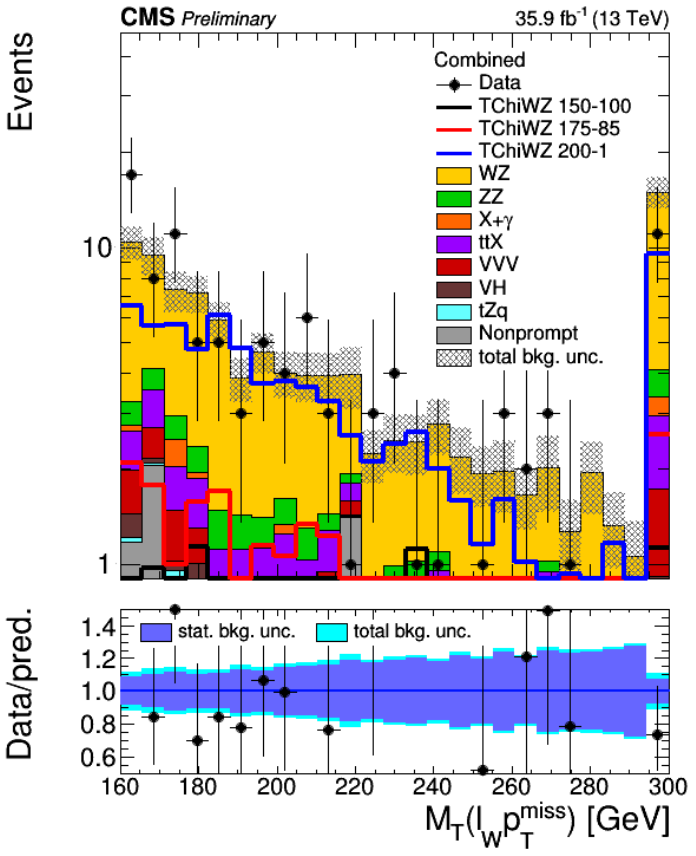

Figura 30: Predicción y datos para el reetiquetado en la región $M_T > 160$ GeV

	N	σ_N
TchiWZ (150-100)	12	2
TchiWZ (175-85)	21	2
TchiWZ (200-1)	79	3
WZ	72	8
ZZ	6.0	0.6
x+γ	1.6	0.8
ttX	10	2
VVV	6	3
VH	2.4	0.9
tZq	0.4	0.2
Nonprompt	14	5

Cuadro 9: Número de eventos y desviación estándar para $M_T > 160$ GeV y para el etiquetado original

	N	σ_N
TchiWZ (150-100)	12	2
TchiWZ (175-85)	21	2
TchiWZ (200-1)	79	3
WZ	68	7
ZZ	5.8	0.5
x+γ	1.5	0.8
ttX	9	1
VVV	6	3
VH	2.5	0.9
tZq	1.4	0.2
Nonprompt	12	4

Cuadro 10: Número de eventos y desviación estándar para $M_T > 160\,GeV$ y para el reetiquetado

Por los cuadros 9 y 10 observamos una reducción del fondo total, que era lo esperado.

	S	σ_S	B	σ_B	D	σ_D	S/B	$\sigma_{S/B}$
Original	12	2	110	10	100	10	0.10	0.04
Reetiquetado	12	2	105	9	100	10	0.11	0.04

Cuadro 11: Señal, fondo, datos y discriminante señal/fondo con sus incertidumbres para $M_T > 160\,GeV$

En el cuadro 11 se observa un incremento del 10 % en el discriminante señal/fondo, que era nuestro objetivo principal.

Debido a esta reducción de fondo, si los datos no se han reducido de la misma manera esperamos un aumento de la intensidad de señal y una disminución de los límites superiores.

	μ	σ_μ	Límite superior	Límite superior esperado
TChiWZ (150-100)	-0.6	1.0	1.4	2.5
TChiWZ (175-85)	-0.3	0.7	0.9	1.3
TChiWZ (200-1)	-0.09	0.2	0.3	0.4

Cuadro 12: Intensidad de señal, desviación estándar de esta, límite superior y límite superior esperado para $M_T > 160\,GeV$ y para el etiquetado original

	μ	σ_μ	Límite superior	Límite superior esperado
TChiWZ (150-100)	-0.6	1.0	1.4	2.3
TChiWZ (175-85)	-0.3	0.6	0.9	1.3
TChiWZ (200-1)	-0.09	0.2	0.3	0.3

Cuadro 13: Intensidad de señal observada, desviación estándar de esta, límite superior (UL) y límite superior esperado para $M_T > 160\,GeV$ y para el reetiquetado

Obtenemos una reducción significativa de los límites superiores esperados debida a la reducción de fondo, como podemos comprobar en los cuadros 12 y 13.

	σ_{theo} (fb)	$\sigma_{\sigma_{theo}}$ (fb)	σ_{obs} (fb)	$\sigma_{\sigma_{obs}}$ (fb)	UL(σ_{obs}) (95 % CL)
TChiWZ (150-100)	5180.86	253.22	-3000	6000	9000
TChiWZ (175-85)	2953.28	154.39	-1000	2000	3000
TChiWZ (200-1)	1807.39	101.32	-200	300	500

Cuadro 14: Secciones eficaces teórica y observada, sus desviaciones estándar y el UL de la sección eficaz observada para la region $M_T > 160$ GeV y el etiquetado original

	σ_{theo} (fb)	$\sigma_{\sigma_{theo}}$ (fb)	σ_{obs} (fb)	$\sigma_{\sigma_{obs}}$ (fb)	UL(σ_{obs}) (95 % CL)
TChiWZ (150-100)	5180.86	253.22	-3000	6000	9000
TChiWZ (175-85)	2953.28	154.39	-1000	2000	3000
TChiWZ (200-1)	1807.39	101.32	-200	300	500

Cuadro 15: Secciones eficaces teórica y observada, sus desviaciones estándar y el UL de la sección eficaz observada para la región $M_T > 160$ GeV y el reetiquetado

En lo respectivo a la σ_{obs} y la precisión de su medida (en cuadros 14 y 15) no conseguimos hacer mejoras significativas.

En definitiva, en esta región el reetiquetado ha aumentado la relación señal/fondo pero no podemos decir que nuestra medida mejore.

Estudio de la región del espacio de fases restringida por $M_{ll} < 50$ GeV

En las figuras 31 y 32 se muestran las predicciones para los dos etiquetados con fin comparativo, para el corte cinemático $M_{ll} < 50$ GeV.

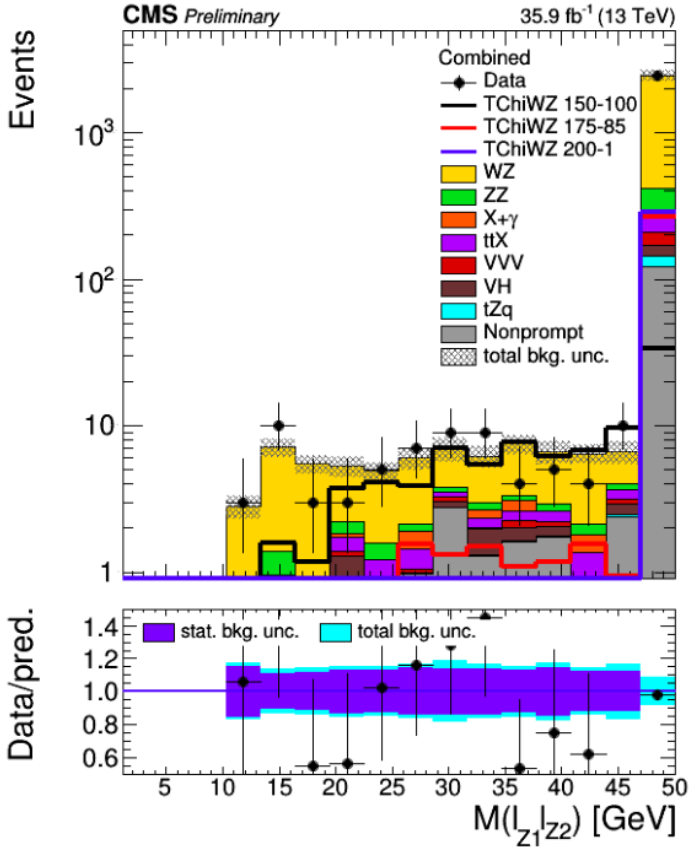

Figura 31: Predicción y datos para el etiquetado original en la región $M_{ll} < 50$ GeV

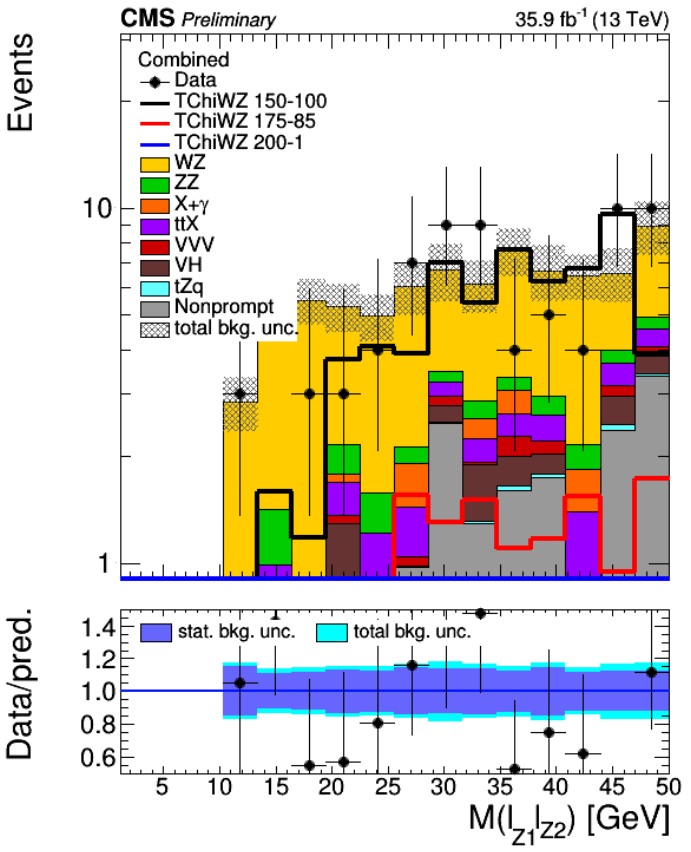

Figura 32: Predicción y datos para el reetiquetado en la región $M_{ll} < 50$ GeV

	N	σ_N
TchiWZ (150-100)	62	5
TchiWZ (175-85)	12	2
TchiWZ (200-1)	2.5	0.5
WZ	48	5
ZZ	4.1	0.4
x+γ	1	1
ttX	4.4	0.7
VVV	1.7	0.9
VH	4	1
tZq	0.3	0.1
Nonprompt	16	5

Cuadro 16: Número de eventos y desviación estándar para $M_{ll} < 50\,GeV$ y para el etiquetado original

	N	σ_N
TchiWZ (150-100)	62	5
TchiWZ (175-85)	12	2
TchiWZ (200-1)	2.5	0.5
WZ	48	5
ZZ	4.1	0.4
x+γ	2	1
ttX	4.4	0.7
VVV	1.7	0.9
VH	4	1
tZq	0.3	0.1
Nonprompt	16	5

Cuadro 17: Número de eventos y desviación estándar para $M_{ll} < 50$ GeV y para el reetiquetado

En los cuadros 16 y 17 vemos que prácticamente no hay variación del fondo total al aplicar el reetiquetado si comparamos con el etiquetado original.

	S	σ_S	B	σ_B	D	σ_D	S/B	$\sigma_{S/B}$
Original	62	5	80	8	82	9	0.8	0.1
Reetiquetado	62	5	80	8	81	9	0.8	0.1

Cuadro 18: Señal, fondo, datos y discriminante señal/fondo con sus incertidumbres para $M_{ll} < 50$ GeV

Las *yields* proporcionadas son prácticamente idénticas, al igual que el discriminante S/B. El reetiquetado no afecta de manera relevante a esta región.

	μ	σ_μ	Límite superior	Límite superior esperado
TChiWZ (150-100)	0.02	0.2	0.4	0.38
TChiWZ (175-85)	0.1	1	2	1.9
TChiWZ (200-1)	0.6	5	10	10

Cuadro 19: Intensidad de señal observada, desviación estándar de esta, límite superior y límite superior esperado para $M_{ll} < 50$ GeV y para el etiquetado original

	μ	σ_μ	Límite superior	Límite superior esperado
TChiWZ (150-100)	0	0.2	0.4	0.38
TChiWZ (175-85)	0	1	2	1.9
TChiWZ (200-1)	0	5	10	10

Cuadro 20: Intensidad de señal observada, desviación estándar de esta, límite superior (UL) y límite superior esperado para $M_{ll} < 50$ GeV y para el reetiquetado

En lo respectivo a la intensidad de señal observada y los límites superiores para el etiquetado original, si comparamos los cuadros 19 y 20 se observa un cambio drástico en la intensidad de señal observada, ya que se anula para los tres puntos de señal. Esto ocurre por que el número de datos y el número de eventos de fondo predichos son muy similares y una pequeña variación de fondo hace que la intensidad de señal observada para puntos con pocos eventos predichos en la región sufra variaciones grandes.

Pasemos a comparar la sección eficaz observada.

	σ_{theo} (fb)	$\sigma_{\sigma_{theo}}$ (fb)	σ_{obs} (fb)	$\sigma_{\sigma_{obs}}$ (fb)	UL(σ_{obs}) (95 % CL)
TChiWZ (150-100)	5180.86	253.22	100	1000	2000
TChiWZ (175-85)	2953.28	154.39	300	3000	6000
TChiWZ (200-1)	1807.39	101.32	1000	9000	19000

Cuadro 21: Secciones eficaces teórica y observada, sus desviaciones estándar y el UL de la sección eficaz observada para la región $M_{ll} < 50$ GeV y el etiquetado original

	σ_{theo} (fb)	$\sigma_{\sigma_{theo}}$ (fb)	σ_{obs} (fb)	$\sigma_{\sigma_{obs}}$ (fb)	UL(σ_{obs}) (95 % CL)
TChiWZ (150-100)	5180.86	253.22	0	1000	2000
TChiWZ (175-85)	2953.28	154.39	0	3000	6000
TChiWZ (200-1)	1807.39	101.32	0	9000	18000

Cuadro 22: Secciones eficaces teórica y observada, sus desviaciones estándar y el UL de la sección eficaz observada para la región $M_{ll} < 50\,GeV$ y el reetiquetado

En los cuadros 21 y 22 observamos que aplicar el reetiquetado nos proporciona una mejoría relevante en la medida de la sección eficaz.

7. Conclusiones

Hemos estudiado la producción de procesos SUSY en el LHC, para un estado final de tres leptones ligeros. Para llevar a cabo este análisis hemos definido tres regiones del espacio de fases del proceso que consideramos más significativas y hemos visto cómo en alguna de ellas el fondo predicho estaba fuertemente influenciado por errores en la asignación de la procedencia de los leptones del estado final.

Partíamos de una tasa de error en el etiquetado de leptones del 2,79 %. Para intentar reducir este error hemos entenado un clasificador binario (tenemos eventos 'mal clasificados' y 'bien clasificados') por red neuronal, y hemos logramos rebajar la tasa de error a un 2,56 %, esto es, conseguimos una mejora relativa del 8,2 % respecto al etiquetado original.

Posteriormente hemos aplicado el reetiquetado con menor tasa de error para realizar una nueva predicción de fondo, y al comparar con el etiquetado original hemos comprobado como la relación señal/fondo mejoraba cerca del 10 % en la región $M_T > 160GeV$, mientras que en la región del espacio de fases completo y la región $M_{ll} < 50$ GeV esta relación señal/fondo tras el reetiquetado era similar. La intensidad de señal observada, su límite superior y la sección eficaz observada así como su límite superior no presentan variaciones significativas tras la aplicación del reetiquetado. La conclusión principal que sacamos de todo este análisis es que se ve compatibilidad con sección eficaz nula.

Como extensiones a este estudio, cabe la posibilidad de variación de algunos hiperparámetros y variables de entrada o el diseño de una función de pérdida específica para el problema de tal manera que se puedan mejorar los resultados obtenidos. Además, este estudio podría extenderse a otros cortes cinemáticos, además de poder considerarse otros puntos de señal.

Referencias

[1] The CMS Collaboration *Search for electroweak production of charginos and neutralinos in multilepton final states in proton-proton collisions at* $\sqrt{s} = 13TeV$ CMS-SUS-16-039 (2018).

[2] The CMS Collaboration *Combined search for electroweak production of charginos and neutralinos in proton-proton collisions at* $\sqrt{s} = 13TeV$ CMS-SUS-17-004 (2018).

[3] Philip Bechtle, Tilman Plehn, Christian Sander *The Status of Supersymmetry after the LHC Run 1* arXiv:1506.03091 (2015).

[4] The CMS Collaboration *The CMS experiment at the CERN LHC* JINST 3 S08004 (2008).

[5] The CMS Collaboration *Particle-flow reconstruction and global event description with the CMS detector* arXiv:1706.04965 (2017).

[6] The CMS Collaboration *Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET* CMS-PAS-PFT-09-001 (2009).

[7] The CMS Collaboration *The CMS Particle Flow Algortihm* arXiv:1401.8155 (2014).

[8] Charu C. Aggarwal *Neural Networks and Deep Learning* (2018).

[9] Sandro Skansi *Introduction to Deep Learning* (2018).

[10] Ian J. R. Aitchison *Supersymmetry and the MSSM: An Elementary Introduction* arXiv:hep-ph/0505105 (2005).

[11] F. Rosenblatt *The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain* Psychological Review Vol. 65, No. 6 (1958).

[12] LISA lab, University of Montreal *Deep Learning Tutorial* Release 0.1 (2015).

[13] Laura Igual, Santi Seguí *Introduction to Data Science* (2017).

[14] Glen Cowan, Kyle Cranmer, Eilam Gross, Ofer Vittels *Asymptotic formulae for likelihood-based tests of new physics* arXiv:1007.1727 (2013).

[15] Michael H. Seymour, Marilyn Marx *Monte Carlo Event Generators* arXiv:1304.6677 (2013).