Universidad de Oviedo

Facultad de Ciencias
Grado en física

Trabajo fin de grado

Estudio del etiquetado de jets con técnicas de aprendizaje automático en el experimento CMS del LHC

Autor:

Rocío Coro Argüelles

Directores:

Javier Fernández Menéndez

Andrea Trapote Fernández

Oviedo, 2024

Índice

Introducción

Los científicos, a lo largo de la historia, se han preguntado constantemente de qué está hecha la materia que nos rodea. A finales del siglo XIX tiene lugar un punto de inflexión, provocado por el descubrimiento del electrón, desbancando al átomo como la unidad más pequeña de la materia. Durante la primera mitad del siglo XX, se suceden múltiples descubrimientos de diferentes partículas, dando paso al nacimiento de la física de partículas, en paralelo con otros nuevos campos, como la física cuántica. A partir de estos descubrimientos, llevados a cabo en el último siglo, se ha revelado que la materia está formada por partículas elementales, que interaccionan entre sí mediante cuatro fuerzas fundamentales. La teoría que unifica la estructura de la materia recibe el nombre de Modelo Estándar.

Hoy en día, la física de partículas se enfrenta a nuevos desafíos, entre los que se encuentra la unificación de la gravedad en el Modelo Estándar, la naturaleza de la materia oscura o el desequilibrio entre materia y anti-materia. Para su estudio, existen diversos aceleradores de partículas repartidos por todo el mundo. El Gran Colisionador de Hadrones (LHC), situado en la Organización Europea para la Investigaión Nuclear (CERN), es, en la actualidad, el acelerador de partículas más potente del mundo. En particular, este trabajo se centrará en uno de sus experimentos, conocido como CMS.

En el LHC, se producen colisiones entre haces de protones. Tras estas colisiones, se originan diversas partículas, que posteriormente se agrupan con otras, para formar lo que conocemos como jets. Así, clasificar el origen de los jets proporciona información clave sobre lo que ocurrió tras la colisión, permitiendo una comprensión más profunda del comportamiento de las partículas elementales. Además, en las colisiones del LHC, las partículas originadas con mayor frecuencia son los quarks y los gluones. Identificarlos correctamente permite distinguir el resto de partículas producidas en el mismo experimento, facilitando la observación de nuevos descubrimientos. En concreto, este trabajo buscará etiquetar de forma precisa los jets originados por gluones y quarks.

Por otro lado, nos encontramos ante el *boom* de la inteligencia artificial y, en particular, del aprendizaje automático. Estas técnicas han supuesto una revolución en el tratamiento y análisis

de datos. En el LHC, se genera una gran cantidad de información, que posteriormente debe ser estudiada. La implementación de técnicas de aprendizaje automático en física de partículas, ha permitido mejorar el estudio de los datos obtenidos en los experimentos. En este trabajo, se utilizarán redes neuronales profundas, un modelo dentro del aprendizaje automático cuyo mecanismo se basa en las conexiones neuronales. Dado que el cerebro es la mayor máquina de procesamiento de información conocida, las redes neuronales profundas buscan emular su funcionamiento de manera artificial, mediante la recreación de conexiones neuronales.

En resumen, este trabajo se basará en la implementación de redes neuronales profundas para clasificar el origen de jets, detectados en el experimento CMS del LHC, buscando mejorar los resultados alcanzados mediante observables físicos. Más aún, el trabajo se distribuye en seis capítulos. Los dos primeros presentan la teoría del Modelo Estándar y el experimento CMS del LHC, explicando su funcionamiento y estructura. A continuación, en el capítulo tres se explorará la física de los jets y sus propiedades. Seguidamente, en el capítulo cuatro se presentará el aprendizaje automático y, en concreto, las redes neuronales profundas. En el quinto capítulo, exploraremos la muestra de datos que se utilizará, abordando su simulación y estructura. Luego, en el sexto capítulo, se expondrán los resultados derivados del análisis de estos datos. Finalmente, el trabajo concluye con las reflexiones y conclusiones correspondientes.

Capítulo 1

Modelo Estándar

El Modelo Estándar es una teoría que explica la estructura fundamental de la materia. Es el resultado de la unificación de numerosos descubrimientos que han llevado a cabo diferentes científicos desde principios del siglo XX.

El Modelo Estándar, establece que toda la materia está formada por unas partículas elementales, dominadas por cuatro fuerzas fundamentales: la electromagnética, la gravitatoria, la débil y la fuerte. Esta teoría asegura que las interacciones entre partículas, están descritas mediante el intercambio de otras. La única fuerza excluída de este Modelo es la gravitatoria, ya que aún no se ha encontrado ninguna partícula que justifique esta interacción.

Esta teoría ha conseguido explicar la mayoría de los experimentos realizados e incluso preveer ciertos fenómenos. Como consecuencia, se considera una teoría física demostrada. [1]

1.1. Partículas elementales

Las partículas elementales constituyen la materia. Podemos agruparlas en dos grandes grupos: los fermiones y los bosones. La principal diferencia entre ambos conjuntos es el espín.

1.1.1. Fermiones

Los fermiones son partículas elementales que se caracterizan por tener un espín semientero. Por tanto, siguen la estadística de Fermi-Dirac, que asegura que la energía de sus estados cuánticos toma únicamente valores discretos.

Dentro de los fermiones, diferenciamos entre leptones y quarks, en función de las intercacciones fundamentales que sufren y su carga eléctrica y de color. Cada grupo está formado por

seis partículas, ordenadas en pares o generaciones. Las partículas más estables y ligeras se encontrarán en la primera generación, mientras que las más pesadas e inestables pertencerán a la tercera. Se desconoce la razón por la que existen únicamente tres generaciones, pero hay evidencia científica de que no existe ninguna más.

Leptones

Los leptones son una clase de fermiones caracterizados por no tener carga de color. Como consecuencia, no sienten la interacción fuerte.

En la primera generación de leptones se encuentran el electrón e^- y el electrón-neutrino ν_{e^-}. El primero tiene carga eléctrica -1 y una masa de 0.511MeV mientras que el ν_{e^-} no tiene carga eléctrica y su masa aún no ha sido determinada, únicamente disponemos de una cota superior de $10^{-6} MeV$.

En la segunda y tercera generación de leptones, encontramos copias de estas partículas que difieren en masa de las anteriores pero que poseen las mismas propiedades, es decir, las mismas interacciones elementales.

En el caso del electrón, su análogo en la segunda generación será el muhón μ^-, que tiene una masa aproximada de $m_{\mu^-} \approx 200 m_{e^-}$. El leptón tau τ^- será el correspondiente en la tercera generación, con una masa de aproximadamente $m_{\tau^-} \approx 3500\ m_{e^-}$. Las interacciones fundamentales que sufren estas tres partículas son la electromágnetica y la débil.

Por otro lado, para el electrón-neutrino, su semejante en la segunda generación será el leptón muhón-neutrino μ_{e^-} y en la tercera, el tau-neutrino τ_{e^-}. Sus masas son tan pequeñas que hasta el momento no ha sido posible obtener un valor preciso. Al no tener carga eléctrica, estas partículas estarán sometidas únicamente a la interacción débil.

Quarks

Los quarks consituyen el segundo tipo de fermiones, caracterizados por tener una carga eléctrica fraccionaria y carga de color, por lo que sienten la interacción fuerte.

Existen seis tipos de quarks diferentes, dos por cada generación. En la primera generación encontramos los quarks $up(u)$ y $down(d)$. Se caracterizan por tener las energías más bajas y una masa entorno a los 3-5 MeV. A continuación, en la segunda generación, se encuentran los quarks $charm(c)$ y $strange(s)$, con masas de 1275 y 95 MeV respectivamente. Por último, los quarks de tercera generación serán los quarks $botom(b)$ y $top(t)$, siendo los más pesados, con masas de

4.18 y 173 GeV respectivamente. De este modo, llamaremos quarks ligeros, a los quarks *u, d y s*.

Los quarks de tipo *u, c y t* tienen carga eléctrica 2/3 mientras que la del resto tiene un valor de -1/3. Una característica distinguida de los quarks, es que no se observan nunca como partículas libres, sino como estados ligados que reciben el nombre de hadrones. Esto se debe al confinamiento cuántico, que se explicará más adelante en este capítulo.

Para cada uno de los fermiones mencionados existe una antipartícula con la misma masa y carga opuesta, tanto eléctrica como de color, en el caso de los quarks. Se denotarán l^+ para los leptones y $\bar{q}$ en el caso de los antiquarks.

1.1.2. Bosones

Los bosones son partículas elementales que se caracterizan por tener un espín entero. Así, se rigen por la estadística de Bose-Einstein.

Como se ha comentado anteriormente, el Modelo Estándar describe las interacciones entre partículas mediante el intercambio de otras partículas. Dichas partículas mediadoras son los bosones, de forma que cada tipo de interacción está mediada por un tipo de bosón. Así, los fermiones intercambian cantidades discretas de energía entre ellos, mediante el intercambio de bosones

En el caso de la interacción electromagnética, la partícula mediadora es el fotón γ. Esta partícula tiene espín 1 y no posee ni carga eléctrica ni masa.

Por otro lado, la interacción fuerte es mediada por el gluón g. Tiene espín 1 y carece de carga eléctrica y masa. Sin embargo, posee carga de color, asociada a esta interacción, lo que le permite interactuar consigo mismo.

La interacción débil está asociada a los bosones W^+, W^- y Z^0. Los dos primeros tienen una masa de 80.4GeV y una carga eléctrica elemental, positiva o negativa. Además, ambos son respectivamente la antipartícula del otro. El bosón Z^0 tiene una masa de 91.2 GeV y es eléctricamente neutro. Además, los tres tienen espín 1.

Por último está el bosón de Higgs. Este último, no está asociado a ninguna interacción, sino que se puede interpretar como una excitación del campo de Higgs, campo cuántico responsable de que las partículas elementales adquieran masa. Carece de carga eléctrica y de color, y su espín es 0.

En la Figura 1.1 se muestra un esquema de las partículas elementales agrupadas según el grupo al que pertenezcan.

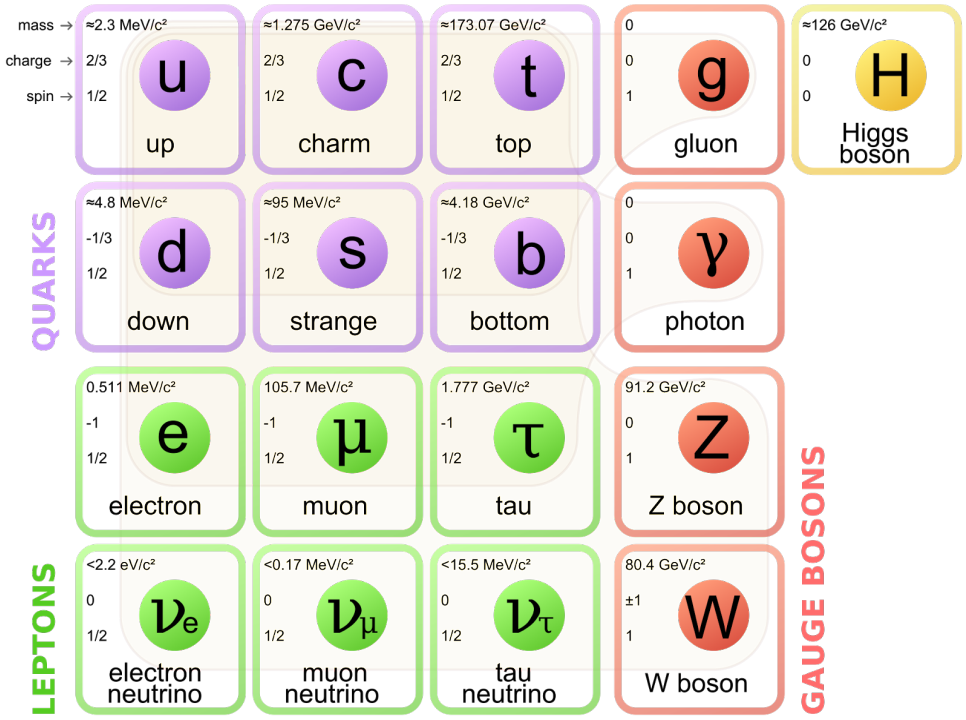


Figura 1.1: En esta figura se muestra un esquema de las partículas elementales, agrupadas según sean quarks, leptones o bosones. Además, la primera columna de fermiones se corresponde a la primera generación, y así de forma consecutiva.[2]

1.2. Interacciones fundamentales

Como ya se ha comentado anteriormente, existen cuatro tipos de interacciones fundamentales: la interacción electromágnetica, la fuerte, la débil y la gravitatoria. Sin embargo, el Modelo Estándar no ha sido capaz de unificar esta última con las anteriores. La fuerza gravitatoria es la más débil de las intercciones fundamentales. Como en los aceleradores de partículas se trabaja con escalas subatómicas, esta fuerza puede considerarse despreciable. Como consecuencia, no se profundizará más sobre la fuerza gravitatoria.

Las interacciones entre partículas se darán con diferente frecuencia, en función de la escala de energía en la que se encuentren. Para poder comparar las tres interacciones, se define la constante de acoplamiento g, que mide la fuerza relativa de cada interacción. En este caso, se utilizará la constante adimensional α, de forma que $\alpha \propto g^2$.

A continuación, se procederá a explicar cada interacción de forma más detallada.

1.2.1. Interacción electromagnética

La interacción electromagnética tiene lugar entre partículas cargadas eléctricamente. Si la carga de estas partículas es del mismo signo, la interacción será repulsiva y, en caso contrario, será atractiva.

La partícula mediadora de esta interacción es el fotón. Al carecer de masa, el rango de dicha interacción es infinito.

La constante α de la interacción electromagnética, viene dada por la constante de estructura fina para bajas energías, $\alpha_{EM} \approx 1/137$. De esta forma, la fuerza de la interacción electromagnética aumenta con la energía y desciende con la distancia.

1.2.2. Interacción fuerte

La interacción fuerte se produce entre partículas que poseen carga de color, es decir, los quarks. La partícula mediadora de esta interacción es el gluón g, que también posee carga de color. Como consecuencia, puede interaccionar consigo mismo. La interacción fuerte está descrita mediante la cromodinámica cuántica (QCD), una teoría cuántica de campos que explica, por tanto, las interacciones entre quarks y gluones.

Una característica fundamental de esta interacción es que no se ha observado experimentalmente ningún quark libre. Este fenómeno se asocia al **confinamiento cuántico**, que establece que las partículas con carga de color no pueden aislarse, y se observan junto a otras, de modo que la combinación resultante, tenga carga de color neta nula. Así, surge la existencia de los hadrones, que se explicarán más adelante. Este comportamiento se debe a que la fuerza fuerte, aumenta con la distancia a un ritmo de 1GeV/fm. De esta forma, si consideramos dos quarks separándose, su interacción puede entenderse como un intercambio de gluones que, a su vez, interactuarán entre sí de forma atractiva, limitando las líneas de campo a un cilindro entre los quarks. En la Figura 1.2 se muestra un esquema de cómo sería está interacción. Podemos aproximar el potencial entre los quarks como:

$$V(r) \approx \kappa r \tag{1.1}$$

donde κ tiene un valor aproximado de 1GeV/fm. Esto implica que la fuerza entre dos quarks libres sería del orden de 10^5N, independientemente de su separación. Como consecuencia, si existieran dos partículas libres con carga de color en el Universo, la energía acumulada en el campo de gluones sería enorme. Dada la tendencia de la naturaleza de converger a la menor energía posible, no podemos encontrar cargas de color libres en el Universo.

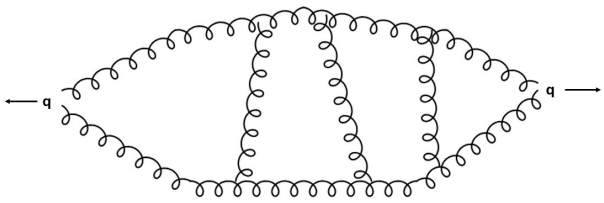


Figura 1.2: Esquema de la interacción que se produciría entre dos quarks libres separándose, en el caso de que pudiera darse esta situación.

Al igual que el fotón, el gluón carece de masa, lo que haría esperar que el rango de interacción de esta fuerza sea infinito. Sin embargo, se ha observado que es del orden de 1fm, debido al confinamiento cuántico.

Por último, la constante de acoplamiento de la interacción fuerte tiene un valor $\alpha_S \approx 1$, siendo el valor más elevado entre las fuerzas fundamentales. La fuerza de esta interacción, disminuye con la energía debido a la libertad asintótica, que establece que en el límite de distancias cortas o, equivalentemente, de altas energías, los quarks interactuan débilmente entre sí.

1.2.3. Interacción débil

La interacción débil está mediada por los bosones W^+, W^- y Z. Es la fuerza capaz de cambiar el sabor de los quarks. Como consecuencia, una partícula como el protón puede cambiar de carga al emitir un bosón W.

Dada la elevada masas de los bosones mediadores, su rango de interacción es pequeño, del orden de $10^{-18}m$, es decir, despaarece fuera del núcleo de una partícula como el protón. Como su nombre indica, se trata de la más débil de las interacciones, con un valor de $\alpha_W \approx 10^{-6}$.

En la tabla 1.1 se muestra un resumen de las fuerzas fundamentales, indicando tanto su constante α asociada, como su rango de actuación y su partícula mediadora .

1.3. Hadrones

Como se ha comentado anteriormente, el concepto de hadrón surge de la necesidad de los quarks de agruparse de forma que la carga de color neta sea nula. Generalmente, los hadrones están formados por 2 o 3 quarks llamados **de valencia**, que son los responsables de sus propiedades. Además, también contienen un **mar de gluones, quarks y antiquarks** que se aniquilan entre sí. Sin embargo, es imposible distinguir estos últimos de un quark de valencia.

Fuerza	α	**Rango (m)**	**Bosón mediador**
Fuerte	1	10^{-15}	Gluones
Electromagnética	1/137	∞	Fotón
Débil	10^{-6}	10^{-18}	Z, W^+, W^-
Gravitatoria	6 10^{-39}	∞	

Cuadro 1.1: Resumen de las cuatro interacciones fundamentales, con el valor α, proporcional a la constante de acomplamiento g, su rango de actuación y sus partículas mediadoras. Aún no se ha encontrado la partícula mediadora de la fuerza gravitatoria. [3]

Cabe mencionar que el único quark que no forma hadrones es el quark *top*, dado que se desintegra antes.

Según el número de quarks de valencia, distinguiremos entre **bariones** y **mesones**. Los primeros se caracterizan por tener tres de quarks de valencia, mientras que los segundos están formados por dos, un quark y un antiquark.

Entre los bariones más conocidos destacan el protón y el neutrón. El protón es un barión formado por dos quarks *up* y un quark *down* mientras que el neutrón está constituido por un quark *up* y dos quarks *down*. En la Figura 1.3 se muestra un esquema de la estructura de un neutrón y un protón, con sus quarks de valencia y mar de quarks, antiquarks y gluones.

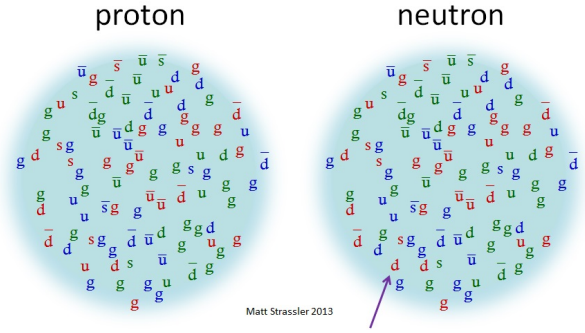


Figura 1.3: Esquema de la estructura interna de un protón y un neutrón, con sus quarks de valencia y el mar de gluones, quarks y antiquarks, que se aniquilan entre sí. [4]

Debido a la finalidad del trabajo, nos centraremos en el estudio de la estructura del protón.

Llamaremos **partones**, a las partículas que constituyen un protón: quarks y gluones. El protón es un barión, con dos quarks *up* y un quark *down* de valencia. Al sumar sus cargas eléctricas, 2/3e y -1/3e respectivamente, obtenemos la carga del protón conocida, +e. Cuando dos protones colisionan, en la interacción participan todos los partones que los constituyen, tanto

los quarks de valencia como todas las partículas del mar. Esto hace que sea tan díficil predecir de forma teórica que ocurrirá tras cada colisión.

Todos los quarks que constituyen el hadrón, aportan una fracción al momento total del mismo. Debido a esto, trabajaremos con distribuciones de la fracción del momento de los partones, *Parton Distributions Functions (PDFs)*, que representan la probabilidad de que cada tipo de partícula lleve una fracción de momento, comprendida entre x y x + δx. Por ejemplo, u(x) representa la probablidad de que un quark u lleve una fracción x del momento total del protón. Las PDFs se obtienen de forma experimental a partir de *scattering* inelástico. En la Figura 1.4 se representan las PDFs para Q^2=10 GeV^2 utilizando el modelo MSTW 2008 next-to-leading-order(NLO).

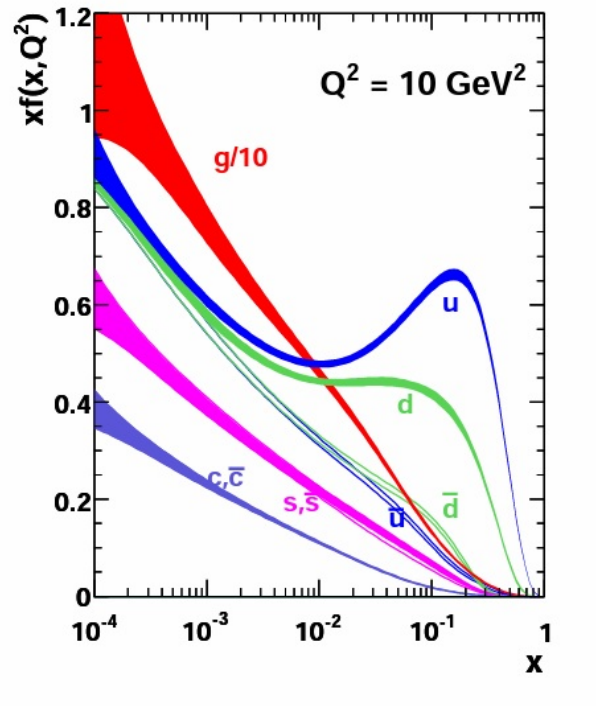


Figura 1.4: Representación gráfica de las PDFs para gluones y quarks con un valor fijo de Q^2=10 GeV^2, utilizando el modelo MSTW 2008 NLO. Cada una de ellas indica la probabilidad de que cada tipo de partícula fundamental lleve una fracción x del momento total del protón. [5]

A lo largo de este trabajo se estudiarán los jets originados en el detector de partículas, que provienen de desintegraciones de hadrones en el propio detector. En el capítulo 3 se profundizará más en este aspecto.

Capítulo 2

El LHC y el experimento CMS

A continuación, se procederá a explicar el Gran Colisionador de Hadrones o *Large Hadron Colllider (LHC)* y, en particular, su experimento CMS. Además, se comentarán diferentes conceptos básicos de la física de aceleradores de partículas, como la sección eficaz o la pseudorapidez, para facilitar la comprensión.

Los aceleradores de partículas son máquinas que utilizan campos electromagnéticos para acelerar partículas cargadas hasta velocidades cercanas a la velocidad de la luz. Pueden ser lineales, en los que las partículas describen un movimiento rectilíneo, o circulares, en los que viajan dos haces, que posteriormente colisionan. Asimismo, los aceleradores son capaces de recrear las condiciones de altas energías del origen del Universo. Su principal aplicación es la investigación en el campo de la física de partículas, con la búsqueda de nuevas partículas o la realización de experimentos para corroborar teorías ya existentes, aunque también se utilizan en otros ámbitos, como la medicina.

2.1. Sección eficaz

En primer lugar, se empezará definiendo el concepto de **sección eficaz** σ. Se trata de una magnitud de gran importancia en las colisiones entre partículas. Se puede interpretar como la probabilidad de que, tras un choque, se produzca una interacción entre las partículas. Se define como:

$$\sigma = \frac{n^{\underline{o}} \text{ interacciones por partícula por unidad de tiempo}}{\text{flujo del haz por unidad de tiempo y superficie}} \tag{2.1}$$

La sección eficaz de un proceso determinado depende del tipo y la energía de las partículas del choque. En el LHC, partículas como los bosones W y Z tienen elevadas secciones eficaces,

es decir, se observarán más a menudo. Por otro lado, el bosón de Higgs tiene una sección eficaz de producción muy pequeña. [6]

La sección eficaz tiene unidades de área y, generalmente, la unidad de medida es el *barn (b)*, tal que 1 b= $10^{-28}m^2$. En una colisión proton-protón del LHC, con una energía $\sqrt{s}$=7 TeV, la sección eficaz total es aproximadamente 100 mb.

2.2. Luminosidad

A continuación, se explicará otra magnitud importante en los aceleradores de partículas: la **luminosidad instantánea** l. Se define como el número de colisiones que pueden producirse en un detector, por unidad de tiempo y de superficie. Por tanto, a mayor luminosidad, mayor cantidad de datos se obtienen del experimento. Para aumentar la luminosidad instantánea del experimento, se puede aumentar el número de partículas en el haz, o estrecharlo para comprimir más las partículas.

Al integrar esta magnitud en un intervalo de tiempo, se obtiene la **luminosidad integrada L**. Esta magnitud indica el número de colisiones que ocurren en dicho intervalo temporal, y se mide en unidades inversas de sección eficaz. [6]

De forma resumida, la luminosidad instantánea mide el número de partículas que atraviesan un área por segundo, mientras que la sección eficaz expresa la probabilidad de que ocurra cierto suceso. Uniendo ambos conceptos, el número de sucesos de determinado proceso por segundo será:

$$N = l\sigma \tag{2.2}$$

En la Figura 2.1 se muestra la evolución de la luminosidad integrada de CMS desde el 2010 al 2023. En concreto, para el 2023 ha sido de 31.4 fb^{-1}.

2.3. Pile up

Un concepto a tener en cuenta en los choques de partículas es el *pile up* o apilamiento. Llamaremos *pile up* a todas las interacciones secundarias que tienen lugar al hacer colisionar dos haces, en nuestro caso de protones.

Cuando dos haces de protones colisionan, se producen numerosos choques protón-protón. Sin embargo, muchos de ellos carecen de interés, ya que no ocurre nada 'raro', es decir, son colisiones

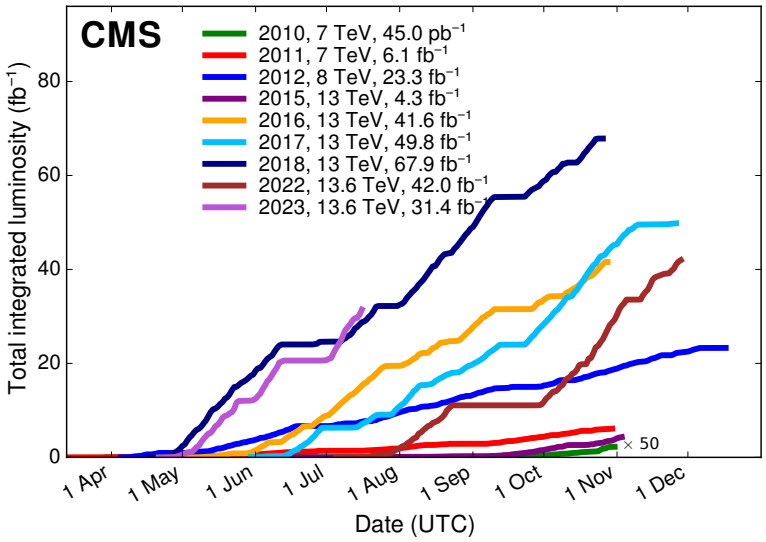


Figura 2.1: Representación gráfica de la evolución de la luminosidad integrada de CMS desde el 2010 hasta el 2023. [7]

que no permitirán observar partículas menos comunes, como el bosón de Higgs. Separar estas interacciones de aquellas relevantes es de vital importancia para el posterior análisis de los datos.

Inicialmente, la técnica para mitigar las interacciones del *pile up* consistía en eliminar las partículas cargadas cuya traza no se originaba en el vértice primario, sino en uno asociado al apilamiento. Esta técnica recibe el nombre de *Charged hadron subtraction* (CHS). Las partículas cargadas suponen el 60 % del total originadas en CMS. Para las partículas neutras, se asume que su energía se distribuye uniformemente por todo el detector, y se sustrae la enería correspondiente.

Actualmente se ha desarrollado una nueva técnica, conocida como *PileUp Per Particle Identification* (PUPPI), que para cada partícula, estima la probabilidad de que haya sido originada por el *pile up*. La técnica PUPPI se basa en que las partículas neutras originadas en la colisión principal, normalmente están alineadas con las partículas cargadas de esta misma colisión, mientras que aquellas origindas por el *pile up*, se distribuyen más uniformemente a lo largo del detector. [8]

La luminosidad instantánea de CMS está en aumento y, por ende, se producirá un mayor número de colisiones en el detector. Como consecuencia, uno de los desafíos actuales es mejorar las técnicas para mitigar las interacciones del *pile up*, ya que su número seguirá aumentando. En la Figura 2.2, se observa la evolución del número de interacciones del *pile up* con los años, en el experimento CMS, en pararelo con el aumento de la sección eficaz.

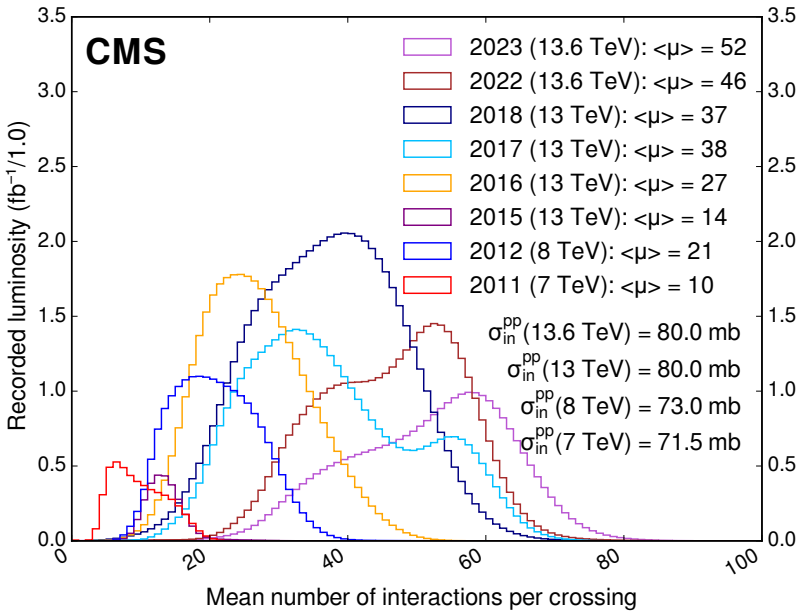


Figura 2.2: Representación gráfica del aumento del número de interacciones de *pile up* por colisión principal en el LHC, a lo largo de los años, debido al aumento de la luminosidad instantánea y la sección eficaz. [9]

2.4. El Gran Colisionador de Hadrones

El Gran Colisionador de Hadrones, o en inglés *Large Hadron Colider* (LHC), es el acelerador de partículas más largo y potente del mundo actualmente, situado en la Organización Europea para la Investigación Nuclear (CERN), en la frontera entre Suiza y Francia, cerca de Ginebra.

El CERN fue creado en 1954 y actualmente cuenta con 22 estados miembros, entre los que se encuentra España. Se trata del mayor laboratorio para la investigación en Física del mundo. Está formado por un conjunto interconectado de aceleradores, que aumentan la velocidad de las partículas de forma gradual. El último eslabón es el conocido Gran Colisionador de Hadrones, que entró en funcionamiento a finales de 2008.

EL LHC un acelerador de partículas de tipo circular, compuesto por un anillo de 27 km de longitud, donde se aceleran protones o iones de plomo para, posteriormente, hacerlos colisionar. Además, el acelerador está situado en el subsuelo a aproximadamente 100 m de profundidad. Dento del acelerador, viajan dos haces de hadrones en sentidos opuestos, en dos tuberias paralelas que se cruzan en cuatro puntos, coincidiendo con los cuatro detectores con los que cuenta: CMS, ATLAS, ALICE y LHCb. Los haces de hadrones serán acelerados hasta velocidades cercanas a la velocidad de la luz, para luego hacerlos colisionar en alguno de estos cuatro detectores. Además, son guiados a través del LHC mediante electroimanes superconductores que se encuen-

tran a una temperatura de $1.9K$. Para alcanzar estas temperaturas se utiliza helio líquido. Los imanes dipolares generan campos opuestos en cada tubería, que origina una fuerza de Lorentz en dirección al centro de la circunferencia, provocando así el movimiento circular de las partículas. Además, se utilizan imanes cuadrupolares para centrar el haz o imanes más potentes para juntar las partículas entre ellas antes de una colisión, incrementando la probabilidad de choque. [10]

Entre sus contribuciones, destaca la observación en 2012 del Bosón de Higgs, gracias a los experimentos ATLAS y CMS, que confirmó las predicciones de Peter Higgs y François Eggert.

El 5 de Julio de 2022 comenzó la tercera serie de toma de datos en el LHC, conocida como *Run 3*. El acelerador estará funcionando durante cuatro años con una energía de colisión record de 13.6TeV.

2.4.1. Cadena de Inyección

Antes de entrar en el LHC, las partículas atraviesan diferentes aceleradores encadenados, que les van aportando velocidad de forma gradual. Este conjunto de aceleradores se conoce como cadena de inyección y está formada por cuatro aceleradores. A parte de acelerar los hadrones para introducirlos en el LHC, algunos de estos aceleradores también cuentan con experimentos propios.

- El primero de ellos es el **Linac4**, un acelerador lineal que impulsa iones H^- hasta una energía de $160MeV$. Tiene una longitud de 86m y está situado 12m bajo la superficie. Este acelerador utiliza cavidades de radiofrecuencia para cargar conductores cilíndricos positiva y negativamente. Alternando estos conductores, los que estén por delante tirarán de las partículas y los de atrás las empujarán, consiguiendo acelerarlas. Además, utiliza imanes cuadrupolares para conseguir estrechar el haz de iones. Este acelerador se empezó a usar en la cadena de inyección del LHC en 2020, y con él se espera aumentar la luminosidad del mismo. Antes de llegar a la siguiente etapa de la cadena de inyección, el Proton Syncroton Booster, los iones H^- pierden sus dos electrones.

- La siguiente etapa es el **Proton Syncroton Booster** (PSB) formado por cuatro anillos sincrotrón superpuestos. Un sincrotrón es un acelerador de partículas circular, que utiliza campos eléctricos y magnéticos variables para impulsar las partículas. El PSB recibe únicamente protones y los acelera hasta una energía de 2GeV.

- La tercera etapa de la cadena de inyección es el **Proton Synchrotron** (PS). Se trata de un sincrotrón con una circunferencia de 628m, que acelera los protones hasta 26GeV. El

PS es el sincrotrón más antiguo del CERN, inagurado en 1959, y capaz de acelerar otros tipos de partículas como partículas α o electrones.

- Por último, las partículas aceleradas atraviesan el **Super Proton Synchrotron** (SPS), formado por 7km de circunferencia, que acelera las partículas hasta 450GeV. Se trata del segundo acelerador más grande del CERN, por detrás del LHC. Proporciona haces acelerados para los experimentos del LHC, NA61/SHINE, NA62 y COMPASS. En 1983 se observaron los bosones W y Z, utilizando el SPS como colisionador de protones y antiprotones.

En la Figura 2.3, se muestra un esquema de la cadena de inyección del LHC, es decir, del recorrido que hacen las partículas para alcanzar la energía necesaria, antes de llegar al LHC.

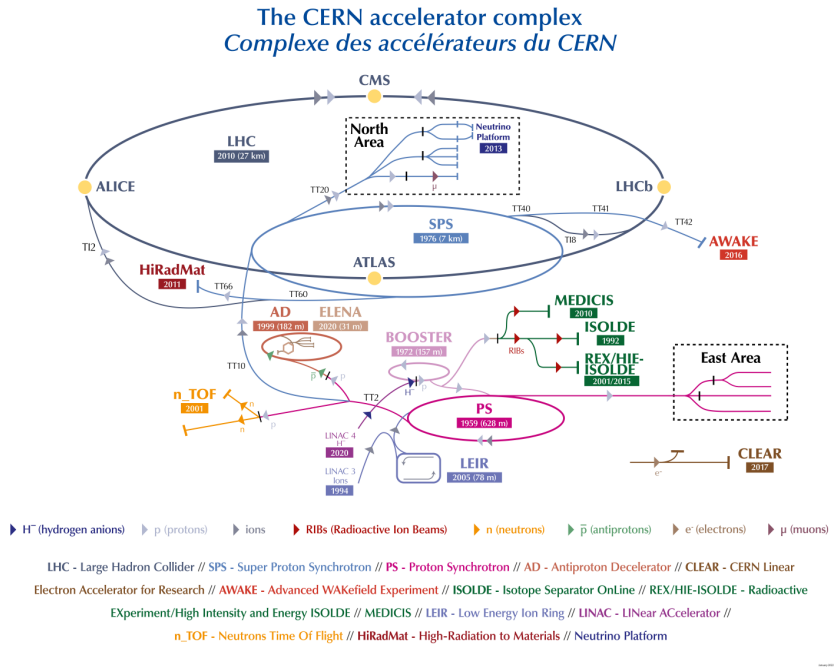


Figura 2.3: Esquema de la cadena de inyección del LHC, formada por cuatro aceleradores de partículas: *Linac4, Proton Syncroton Booster, Proton Synchrotron y Super Proton Synchrotron.* Cuando las partículas atraviesan los aceleradores, van ganando energía de forma gradual, antes de llegar al LHC, con una energía de hasta 7 TeV. [11]

2.4.2. Experimentos

El LHC está formado por cuatro detectores diferentes. Cada uno de ellos tiene asociado un experimento en función de sus características. Además, hay cinco experimentos adicionales, que obtienen la información de los cuatro detectores previos. Los experimentos están llevados a cabo

por colaboraciones entre científicos de diferentes instituciones internacionales. Los experimentos más grandes son ATLAS y CMS. Ambos son detectores de propósito general, pero diseñados de forma independiente, para la confirmación cruzada de nuevos descubrimientos. Por el contrario, los detectores ALICE y LHCb fueron construidos para fenómenos concretos. [12]

- ATLAS: Se trata de un detector de caracter general, utilizado para la investigación de un amplio rango de cuestiones, desde el Bosón de Higgs hasta materia oscura. Tiene los mismos propósitos que CMS, aunque su diseño es muy diferente. Está formado por seis capas de detectores que permiten identificar las partículas producidas en la colisión. Es el detector mas grande del mundo, pesando 7000 toneladas.

- CMS: Al igual que el detector ATLAS, se trata de un experimento de caracter general. Dado que la muestra de este trabajo proviene de este experimento, se profundizará más en la sección 2.5.

- ALICE: Este detector está dedicado al estudio de la física de iones pesados. En particular, estudia una fase de la materia llamada plasma quark-gluon, que consiste en materia con densidades extremas de energía, que interactúa fuertemente. Cuando en el LHC colisionan iones de plomo, se recrean condiciones similares a la situación del Universo tras el Big Bang. El detector ALICE estudia cómo se expande y enfría el plasma quark-gluon, observando cómo progresivamente da lugar a las partículas que constituyen la materia en la actualidad.

- LHCb: El *Large Hadron Collider beauty* estudia la diferencia entre la materia y la antimateria a partir de los quarks b. Al contrario que en los detectores ATLAS o CMS, que rodean toda la colisión, el detector LHCb busca únicamente información de partículas viajando en una determinada dirección tras el choque. Esta formado por varios subdetectores consecutivos, con una longitud en conjunto de 20m.

- LHCf: El *Large Hadron Collider forward* es un experimento que utiliza partículas producidas en las colisiones del LHC como fuente para simular rayos cósmicos, en condiciones de un laboratorio. El objetivo es entender cómo las colisiones dentro del LHC, originan cascadas de partículas, similares a los rayos cósmicos interactuando con los núcleos de la atmósfera.

- TOTEM: Este experimento consiste en el estudio de los protones. En particular, los protones originados tras colisiones protón-protón con un ángulo pequeño, región inaccesible en otros experimentos. Los protones estudiados han sido originados en la colisión del detector CMS.

- MoEDAL-MAPP: Se trata de un experimento cuyo objetivo es la búsqueda del monopolo magnético, es decir, una partícula con carga magnética postiva o negativa, en vez de ambas. También está preparado para la detección de nuevas partículas. Si existen partículas teóricas exóticas, estas crearían un rastro diminuto al atravesar el detector MoEDAL.

- FASER: El *Forward Search Experiment* busca detectar partículas ligeras y con interacciones muy débiles, ya que los detectores principales del LHC no están preparados para su detección. Además, también está diseñado para la detección de neutrinos, hasta ahora imposible por su débil interacción con la materia.

- *SND@LHC*: EL *Scattering Neutrino Detector* está, como su nombre indica, diseñado para la detección de neutrinos. Está formado por un detector de neutrinos, seguido de un dispositivo para medir su energía y detectar muones producidos cuando los neutrinos interactúan con la materia.

2.5. El experimento CMS

Como ya se ha comentado, el *Compact Muon Solenoid*, CMS, es uno de los cuatro detectores del LHC. Se trata de un detector de caracter general, utilizado para estudiar amplios campos de la física, desde el bosón de Higgs hasta materia oscura, aunque está diseñado principalemente para la detección precisa de muones.

El detector está contruido alrededor de un imán solenoide gigante, que genera un campo magnético de hasta $4T$. Tiene forma de bobina cilíndrica, con una longitud de 28.7 m y un diámetro de 15 m, y su peso es de 14000 toneladas. Asimismo, está formado por piezas concétricas cilíndricas que detectan las partículas originadas tras las colisiones, dirigidas en todas direcciones. Estas componentes ayudan a recrear una imagen de la colisión, a partir de las propiedades de las partículas detectadas y su trayectoria. Las piezas concéntricas que conforman el detector se visualizan en la Figura 2.4.

Como curiosidad, es el único detector que no fue contruido *in situ*, sino que se montó por partes, realizadas separadamente en la superficie. El experimento CMS es una de las mayores colaboraciones científicas de la historia, englobando aproximadamente 5500 científicos de 241 instituciones a lo largo del mundo.

A continuación se dará una breve explicación de las diferentes partes de CMS, desde la más interna, cercana a la colisión, hacia el exterior.

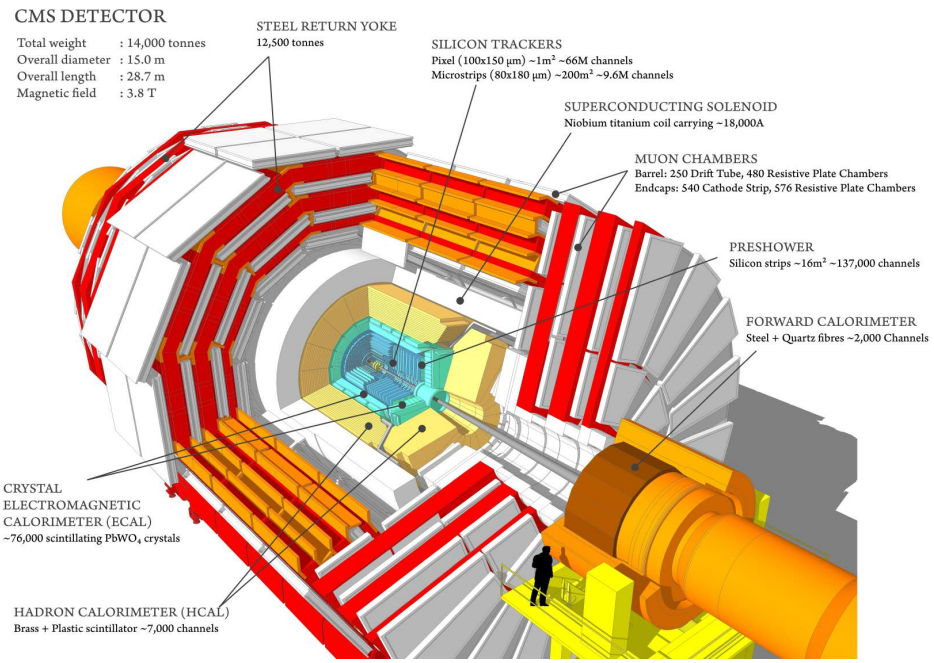


Figura 2.4: Estructura del detector CMS del LHC, formado por capas concentrícas cilíndricas. La capa más interna será el detector de trazas, capaz de medir la trayectoria de las partículas cargadas. A continuación, están los dos calorímetros, que detectan la energía de las partículas que los atraviesan. Por último está el imán del solenoide, que mide el momento de las partículas, y fuera del solenoide, encontramos los detectores de muones.[13]

El detector de trazas

El detector de trazas se encarga de medir e identificar la trayectoria de las partículas cargadas, con una precisión de $10\mu m$. Se trata de la parte más interna de CMS, luego lo atraviesan un gran número de partículas, alrededor de 600 millones de partículas por cm^2/s en la capa más interna. Para reconstruir la trayectoria de una partícula, el detector mide su posición en muchos puntos consecutivos. El detector de trazas es capaz de reconstruir la trayectoria de hadrones, muones y electrones. Además, es capaz de observar trazas originadas por el decaimiento de partículas con poca vida media, como es el caso de los quarks b, y de detectar los vértices secundarios que originan.

Está formado por láminas de silicio dispuestas en capas concéntricas. Podemos dividir este detector en dos: los píxeles, que son la parte más interna, y las microtiras. Los primeros se agrupan en cuatro estructuras concéntricas con forma cilíndrica. Cuando una partícula cargada atraviesa los píxeles, proporciona la suficiente energía como para arrancar los electrones de los átomos de silicio, generando una señal eléctrica pequeña, que posteriormente es amplificada. Las señales se almacenan en chips durante varios microsegundos y después son procesadas. Una vez

conocidos los píxeles atravesados por cierta partícula, se puede reconstruir su trayectoria. Una vez superados los píxeles, las partículas llegan al subdetector de tiras. Este está constituído por 10 capas de detectores de silicio, que alcanzan un radio de 130 cm. Su funcionamiento es análogo al de los píxeles.

Calorímetro electromagnético

A continuación, se encuentran dos calorímetros, utilizados para conocer la energía de las partículas originadas en la colisión. El primero de ellos, el más interno, es el calorímetro electromágnetico, ECAL, que mide la energía de electrones y fotones. Está formado por casi 80000 cristales de tungstato de plomo, de forma que centellea cuando lo atraviesan protones o electrones, produciendo luz en proporción a la energía de la partícula que lo atraviesa. Además, estos cristales producen las lluvias de fotones rápido y bien definidas y, junto a su alta densidad, proporcionan información detallada del recorrido de la partícula cargada. Detrás de cada cristal, se situan fotodetectores, que transforman la luz en una señal eléctrica, que es amplificada y analizada posteriormente.

Cuando un electrón de alta energía llega al calorímetro, interacciona con los núcleos pesados del cristal, excitando los electrones de los átomos, que se relajan rápidamente emitiendo un fotón. Cuando dicho fotón atraviesa la capa de plomo, provoca una lluvia electromagnética, formada por pares electrón-positrón, que los sensores de silicio detectan. De este modo se obtiene una medida de la energía del fotón. Los electrones y fotones pierden su energía y no avanzan más por el detector.

Calorímetro de hadrones

Seguidamente, se encuentra el calorímetro de hadrones, que mide la energía, posición y tiempo de llegada de hadrones, es decir, de partículas constituidas por gluones y quarks. Está formado por una sucesión de capas, alternando un material absorbente y denso con un centellador, que produce luz cada vez que una partícula lo atraviesa. Al igual que en el calorímetro electromagnético, la cantidad de luz emitida se utiliza para medir la energía de la partícula detectada. Cuando un hadrón llega al material absorbente, interacciona con él, produciendo numerosas partículas secundarias, que a su vez interaccionan con las siguientes capas del calorímetro, originando más partículas. Se crea así una cascada de partículas que van atravesando las capas del centellador, que produce luz azul cada vez que una de ellas lo atraviesa. En este calorímetro serán detectados y frenados protones, neutrones, kaones o piones, mientras que los muones y

neutrinos pasarán sin ser identificados.

Además, este calorímetro destaca por ser hermético, construido con capas de forma escalonada para que no haya huecos por los que pudiera escapar una partícula conocida. El objetivo es que atrape todas las partículas producidas a partir del decaimiento los hadrones, y que no serían detectadas en las siguientes capas. De este modo, si observamos un desequilibrio en el momento transverso y la energía, podemos deducir que estamos produciendo partículas "*invisibles*".

Imán del solenoide superconductor

El imán del solenoide superconductor se utiliza para curvar las partículas cargadas que llegan al detector. Se encuentra rodeando a los calorímetros. Para el mismo campo magnético, las partículas cargadas positiva y negativamente se curvan en direcciones opuestas. Además, también permite medir el momento de las partículas, ya que, aquellas con un valor más elevado curvarán su trayectoria en menor medida. Tiene una longitud de 13m y 6m de diámetro, y es capaz de producir un campo magnético de 4T, aunque para alargar su longevidad se utiliza a 3.8T.

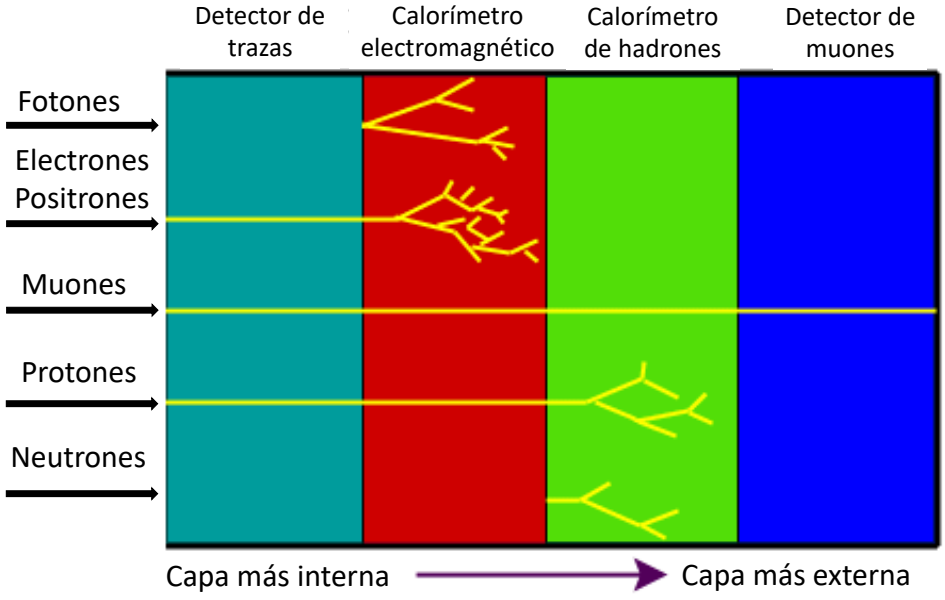


Figura 2.5: Esquema de las partículas que atraviesan el detector CMS, señalando qué subdetectores son capaces de localizarlas. Además, se observa en qué parte del detector cada partícula pierde su energía. Destaca cómo los muones son capaces de atravesar el detector en su totalidad, sin perder su energía. [14]

Detectores de muones

Por último, encontramos los detectores de muones. La detección de estas partículas es una de las principales características de CMS. Los muones son unas partículas muy pesadas, capaces de atravesar metros de materia sin perder mucha energía, por lo que no son detectados ni en el detector de trazas ni en los calorímetros. Las cámaras para su medicción están situadas fuera del solenoide.

Los detectores de muones cuentan con cuatro estaciones, cada una con múltiples capas, donde se registran curvas asociadas a los muones. Estos detectores son capaces de reconstruir su trayectoria, agregando las curvas de cada estación. El campo ejercido sobre los muones será de 2T y en sentido contrario al del solenoide. [15]

La primera estación es el *Muon Drift Tube* (DT), que recubre la parte central del detector. Está formado por un alambre estirado dentro de un volumen de gas. Cuando un muon atraviesa el gas, libera a los electrones de sus átomos, que se desplazan hacia el ánodo siguiendo el campo magnético, para posteriormente ser amplificados y producir un pulso medible. A partir del tiempo que tardan los electrones en alcanzar el cable, se puede determinar la posición del muon. Por otro lado, encontramos el conjunto de cámaras de bandas catódicas (CSC), situadas en los tapones del detector. Estas cámaras están formadas por cables cargados positivamente, cruzados con tiras de cátodo de cobre cargadas negativamente, dentro de un volumen de gas. De forma análoga, cuando pasa un muon, se libera un electrón del gas, que se desplaza hasta el ánodo creando un pulso eléctrico. A continuación, se situan las cámaras de placas resistivas (RPCs), formadas por dos placas paralelas, un ánodo cargado postivamente y un cátodo cargado negativamente, separadas por un gas. Al igual que antes, el paso de un muon produce una señal eléctrica que es recogida. Por último, están los mutiplicadores de electrones de gas (GEMs), que se trata de la última incorporación a los detectores de muones. Están formados por un polímero aislante rodeado de conductores de cobre y un volumen de gas que se ioniza con el paso de un muon, produciendo una señal eléctrica. A partir de todas las señales eléctricas recogidas se reconstruyen las curvas de los muones.

Detectar muones es de vital importancia porque, debido a su elevada masa, son partículas muy probables de originarse en el decaimiento de nuevas partículas pesadas.

Las únicas partículas conocidas que no dejan huella en el detector son los neutrinos, aunque se sabe que tras una colisión protón-protón se producirán muchos de ellos.

A modo de resumen, en la Figura 2.5 proporciona un esquema que indica las partes del detector donde es posible observar cada tipo de partícula. De manera complementaria, la Figura

2.6, muestra un corte transversal del CMS, junto con el recorrido que haría cada una de las partículas, hasta perder toda su energía.

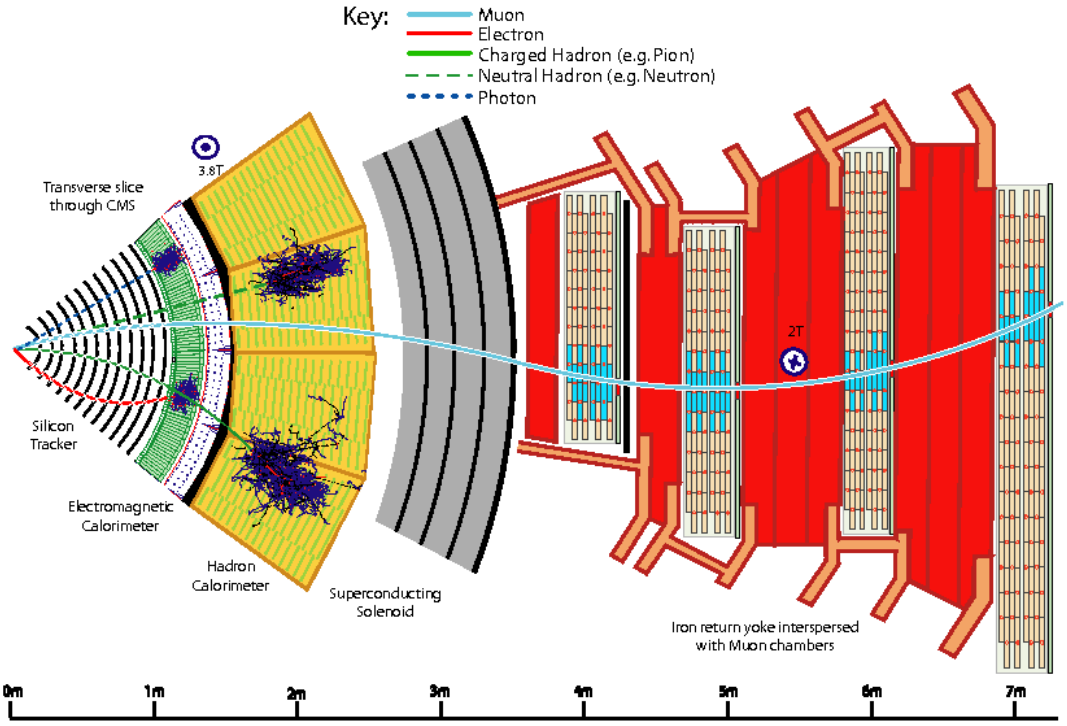


Figura 2.6: Esquema del recorrido de los diferentes tipos de particulas a través de CMS. Se observa como los electrones y fotones pierden toda su energía en el ECAL, y no avanzan más por el detector, mientras que los hadrones serán frenados en el calorímetro de hadrones. Por otro lado, los muones atraviesan todo el detector, hasta los detectores de muones. [16]

2.6. Sistema de referencia

Por último, se va a explicar el sistema de referencia usado en el LHC. La necesidad de definir un sistema de referencia propio surge del hecho de que en el LHC estamos trabajando con partículas relativistas, y buscamos una magnitud invariante de Lorentz que facilite su determinación.

En primer lugar, se va a definir el sistema de coordenadas cartesianas, cuyo origen será el centro de la colisión. El eje Z será paralelo a las direcciones de los dos haces, a lo largo del cilindro del LHC, mientras que el plano X-Y será perpendicular a esta dirección.

A partir de estas coordenadas, definimos unas coordenadas esféricas donde el ángulo azimutal ϕ, es el ángulo entre el eje X e Y, medido desde el eje X, y el ángulo polar θ es aquel que forman el eje X y Z. El primero de ellos toma valores entre $[-\pi, \pi]$ y el segundo entre $[-\pi/2, \pi/2]$.

A partir de este sistema de referencia, se define la coordenada espacial **pseudorapidez (η)** como:

$$\eta = -ln\left(tan\left(\frac{\theta}{2}\right)\right) \tag{2.3}$$

tal que toma valores entre -∞ y ∞. De esta forma si η=0 significa que la dirección de la partícula tras el choque es complemante perpendicular a la dirección del haz, θ=90. Por el contrario, si η=∞, quiere decir que la partícula tras el choque se desplaza en la dirección del haz, θ =0. Por convenio, la pseudorapidez será positiva cuando esté en la dirección del macizo del Jura, que se encuentra en los Alpes suizos.

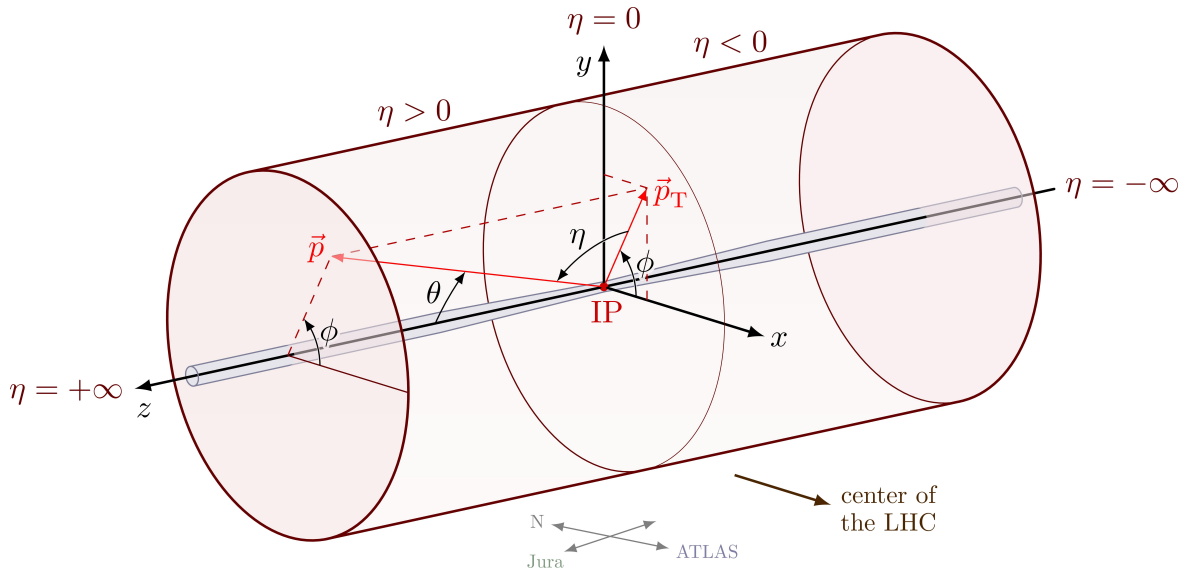


Figura 2.7: Esquema del sistema de referencia del CMS, donde aparecen las coordenadas cartesianas y cilíndricas junto a la pseudorapidez. [17]

El detector CMS permite observar partículas que tras la colisión han salido en una dirección con $|\eta|$ <2.5.

Finalmente, el sistema de coordenadas que se utilizará en el LHC será:

$$p = (p_t, \eta, \phi) \tag{2.4}$$

donde p_t es el momento transverso de la partícula, es decir, la componente del momento en el plano X-Y y, por tanto, $p_t = \sqrt{p_x^2 + p_y^2}$. También se puede expresar en función de la pseudorapidez como:

$$p_t = \frac{|p|}{cosh(\eta)} \tag{2.5}$$

En la Figura 2.7 se esquematiza el sistema de referencia de CMS, junto con las coordenadas cartesianas.

2.7. Trigger y sistema de computación

Cuando el detector CMS funciona a su máxima capacidad, los protones colisionan 40 millones de veces por segundo, en el interior del detector, con únicamente 25 ns ente cada choque. Resulta imposible leer esta cantidad de datos y, de ser posible, sería menos probable que revelasen eventos de interés. Como consecuencia, existe un proceso capaz de seleccionar los datos con un mayor interés potencial, es decir, aquellos que podrían dar lugar a partículas como el bosón de Higgs. Este proceso recibe el nombre de *trigger*. Una vez reducido el volumen de datos, resulta más fácil almacenarlo y analizarlo.

El *trigger* consta de dos niveles: L1 y HLT. El primer nivel es un proceso automático y rápido, que se encarga de buscar partículas con mucha energía o combinaciones inusuales. Selecciona los 100,000 eventos más intersantes por segundo. El segundo nivel se encarga de agrupar la información de las diferentes partes del detector para recrear el suceso y, posteriormente, mandarla a más de 1000 ordenadores. Estos ordenadores revisan la información durante mayor tiempo, relacionando las diferentes medidas. [18]

Aún así, el detector CMS funcionando a pleno rendimiento produce más de 5 petabytes por año. Para almacenar esta información, el LHC utiliza una infraestructura de computación y almacenamiento de datos distribuida, denominada *Worldwide LHC Computing Grid* (WLCG). A través de esta infraestructura, decenas de miles de ordenadores estándar colaboran en todo el mundo para disponer de una gran capacidad de procesamiento, mayor de la que se alcanzaría con un sólo super ordenador. De esta forma, miles de científicos de todo el mundo tienen acceso a los datos.

En el nivel cero del tratamiento de los datos, el CERN reconstruye los eventos tras la colisión y se realiza un control de calidad de los mismos. A continuación, estos datos se envían a grandes centros informáticos en siete paises diferentes, lo que se conoce como el nivel 1. Allí, se vuelven

a reconstruir los sucesos y a cotejar los resultados en búsqueda de patrones. Por último, en el nivel 2 se envian los sucesos más complejos a una serie de instalaciones, para un análisis más específico.

Capítulo 3

Jets

Los jets son agupaciones espaciales de partículas con una vida media larga, producidas por la hadronización de un quark o un gluon. Por tanto, los jets se originan a partir de los hadrones, que a su vez son agrupaciones de quarks con carga de color neta nula. También podemos entender los jets como haces colimados de hadrones. En CMS, los jets están formados principalmente por estas partículas:

- 65 % hadrones cargados.
- 25 % fotones.
- 10 % hadrones neutros.

Asimismo, los jets son muy desordenados, ya que los decaimientos de hadrones inestables pueden producir cientos de partículas muy juntas en el detector, de ahí, que no se analicen partículas individualmente. Sin embargo, las propiedades cinemáticas de los jets se asemejan a aquellas de los partones iniciales que los originan.

A continuación, se explicarán las propiedades de los jets. Luego, se especificarán las características distintivas de los jets originados a partir de quarks ligeros o gluones, en consonancia con el objetivo principal de este trabajo, que es etiquetar estas clases particulares de jets.

3.1. Física de los jets

Como ya se ha comentado anteriormente, en una colisión protón-protón participan todas las componentes del protón, es decir, los gluones y los quarks, de valencia y de mar. Tras la

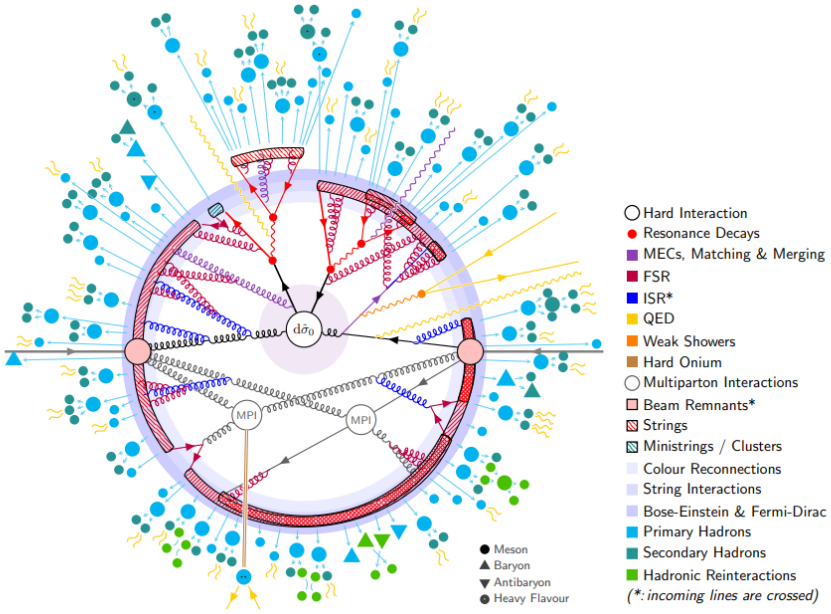


Figura 3.1: Esquema de los procesos que ocurren como consecuencia de una colisión protón-protón. Los partones empiezan a agruparse en hadrones para posteriormente producir jets. Finalmente, los hadrones que forman los jets decaen en partículas más estables. Al mismo tiempo, ocurren procesos subyacentes o *pile up*. [19]

colisión, estos partones se liberan del protón. Sin embargo, de acuerdo con el confinamiento cuántico, no pueden existir en libertad. Como consecuencia, empiezan a agruparse en hadrones o a desintegrarse para posteriormente unirse.

En la Figura 3.1 se muestra un esquema de los procesos que desencadena una colisión protón-protón. En primer lugar, tras el choque, se liberan todos los partones, que empiezan a interaccionar entre sí, produciendo lo que se conoce como cascada de partones. En este momento, los quarks comienzan a separarse entre sí a gran velocidad, aumentado su energía mediante el intercambio de gluones. Una vez que han alcanzado una energía mayor que $2m_q$, estos pueden desintegrarse creando un nuevo par quark-antiquark $\bar{q}q$. Cuando la energía disminuye, se separan en dos pares de hadrones formados por el par quark-antiquark. Este proceso se repite a medida que los quarks se van separando, produciendo numerosos pares $\bar{q}q$. Este mecanismo recibe el nombre de **hadronización**, ya que como resultado se obtienen numerosos hadrones. En la Figura 3.2 se muestra un esquema del proceso de hadronización de dos quarks.

Los partones que, en un primer momento, no se hadronizan, interactúan entre sí dando lugar a otras partículas. En la Figura 3.3, se muestran varios diagramas de Feymann de interacciones que podrían ocurrir en esa situación. Posteriormente, las partículas resultantes se acaban hadronizando también, ya que, en su mayoría, darán lugar a quarks o gluones.

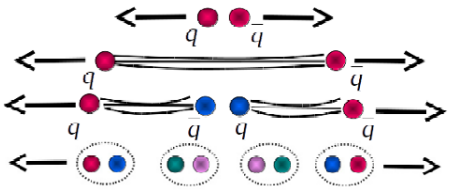


Figura 3.2: Proceso de hadronización de dos quarks. Comienzan alejándose a grandes velocidades, para después dar lugar a pares $\bar{q}q$ y terminar formando numerosos hadrones. [20]

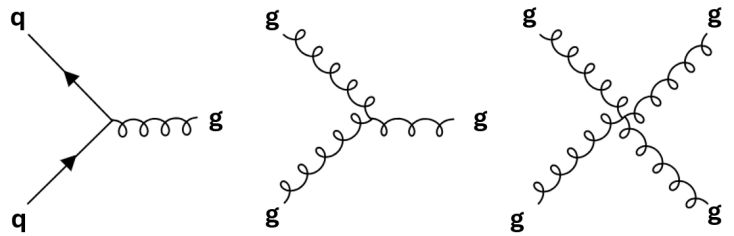


Figura 3.3: Ejemplos de diagramas de Feymann para QCD.

Estos hadrones resultantes se agrupan formando **jets** y se desplazarán en la dirección del par $\bar{q}q$ original. En la Figura 3.4 se observa un esquema del proceso de formación de un jet, desde el partón liberado en la colisión hasta su posterior hadronización para terminar originando un jet.

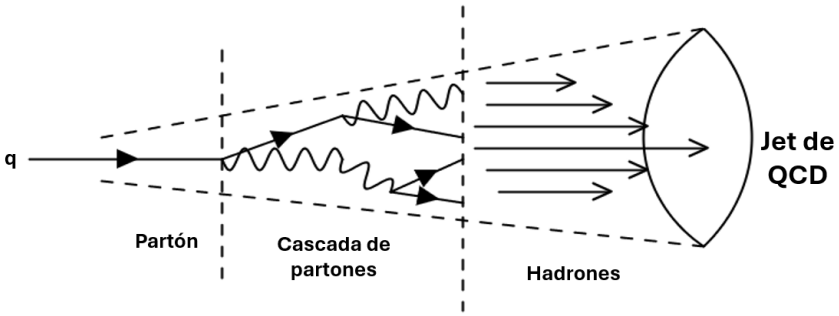


Figura 3.4: Esquema del proceso de formación de un jet. Tras la colisión protón-protón, un partón interactúa con otros en la cascada de partones para posteriormente hadronizarse. Finalmente, varios hadrones forman un jet.

Cuando un jet llega al detector, generalmente no se conoce el sabor del quark o gluón que lo originó. Sin embargo, existen ciertas situaciones que permiten conocer información sobre el origen del jet.

En el caso de los quarks b, al ser los más pesados, su proceso de hadronización dará lugar a un jet con al menos un hadrón que contenga un quark de este sabor. Además, estos hadrones tienen una vida media muy corta, $\tau \approx 10^{-12} s$. Como consecuencia, viajan unos pocos milímetros

antes de decaer. Los hadrones con un quark b serán los únicos que sufran este decaimiento, ya que tienen una vida media más corta y decaen mucho antes.

Se llamará **vértice primario** al punto dónde ocurre la interacción protón-protón, y **vértice secundario** al lugar dónde se produce el decaimiento del hadrón b. Como consecuencia, diferenciar ambos vértices permitirá identificar si el jet se originó a raiz de un quark b. En la Figura 3.5 se muestra un esquema de la distancia que recorre un hadrón b con respecto al vértice primario antes de decaer. Esta distancia viene dada por:

$$d = \beta c \gamma \tau \tag{3.1}$$

donde c es la velocidad de la luz en el vacío, $\beta = \frac{v}{c}$ la velocidad relativa del hadrón respecto a la de la luz, $\gamma = \frac{1}{\sqrt{1-\beta^2}}$ el factor de Lorentz y τ la vida media del hadrón.

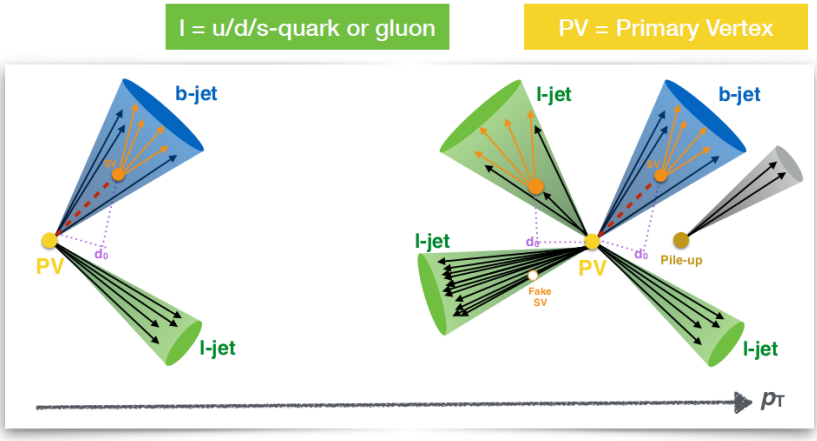


Figura 3.5: Comparación entre la formación de un jet mediante gluones o quarks ligeros en contraste con un quark b. Se observa como este último decae en el conocido como vértice secundario. [21]

Actualmente existen algoritmos capaces de identificar los vértices secundarios y, por tanto, distinguir los jets originados por quarks b.

Sin embargo, este trabajo se centrará en identificar si los jets detectados provienen de quarks ligeros, *up, down y strange* o de gluones. Por tanto, no se profundizará más en la física de los quarks b.

3.2. Jets de gluones y quarks ligeros *u, d y s*

A continuación, dada la finalidad del trabajo, se expondrán las principales diferencias *a priori* entre los jets originados por gluones y por quarks ligeros *u, d y d*.

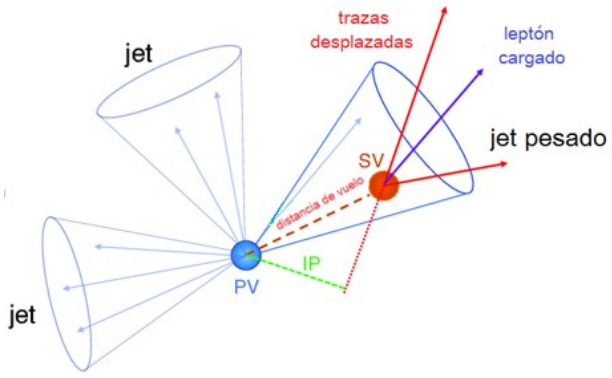


Figura 3.6: Esquema de la distancia que recorre un hadrón b antes de decaer, originando un vértice secundario (SV), con respecto al vértice primario (PV), lugar dónde ocurre la colisión protón-protón principal. [22]

Para diferenciar estos jets de aquellos originados por un quark b se utiliza la definición de vértice secundario ya explicada anteriormente. Si al reconstruir las trazas del jet se observa que convergen en un punto diferente al vértice primario, se concluye que se trata de un vértice secundario y por tanto, de un jet originado por un quark b. Sin embargo, dado que esto se escapa del objetivo del trabajo, no se profundizará más.

Volviendo a los jets originados por quarks ligeros y gluones, se conocen algunas diferencias entre ellos. En el proceso de hadronización, la emisión de gluones es proporcional al factor de color asociado con la interacción del gluón emitido con el emisor. Este factor de color toma el valor C_A=3 cuando el emisor es un gluón y C_F=4/3 en el caso de un quark. Como consecuencia, la multiplicidad de hadrones en un jet originado por un gluón es asintóticamente C_A/C_F=9/4 mayor que en el caso de estar originado por quarks ligeros. Además, los jets originados por un gluon, están menos colimados que aquellos originados por quarks ligeros, es decir, son más amplios. Esto se debe a que la radiación de gluones suaves, aquellos de bajo momento y colineales con la dirección del jet, es mayor para los jets provenientes de gluones. Por último, los quarks producen jets formados por constituyentes *hard*, es decir, que llevan una gran fracción de su momento. [23]

De acuerdo con estas diferencias, se construyen los siguientes observables físicos:

Multiplicidad

La **multiplicidad** de un jet se define como el número de partículas reconstruidas (*PFCandidates*) dentro del jet. Los jets originados por un quark ligero suelen tener menor multiplicidad que aquellos originados por gluones. Además, el ratio de las multiplicidades de los jets provenientes

de quark débiles y gluones converge a

$$\frac{C_A}{C_F} = \frac{9}{4} \tag{3.2}$$

donde C_A y C_F son los factores de color asociado con el acoplamiento del gluon emitido con la partícula emisora. El factor de color indica la probabilidad relativa de que un gluon suave interaccione/se acople con otro gluon. [24]

Anchura

Se puede considerar que los jets tienen forma de cono. Para hablar de anchura de un jet, proyectamos dicho cono sobre el plano $\eta - \phi$, obteniendo una elipse. Normalmente, los jets originados por gluones son más anchos que aquellos originados por quark ligeros, es decir, las trazas correspondientes a las partículas de un jet proveniente de un quark tiene una proyección más circular en el plano $\eta - \phi$. La anchura del jet σ está relacionada con los ejes de la elipse, según la ecuación 3.3, donde σ_1 y σ_2 son los ejes mayor y menor respectivamente. [25]

$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2} \tag{3.3}$$

A menudo se utiliza el eje menor de la elipse en vez de la anchura del jet por simplicidad. Además, ambos ejes de la elipse son función del momento transverso de las componentes del jet.

Función de fragmentación

La función de fragmentación se define como la distribución del momento longitudinal de los hadrones dentro de un jet reconstruido. Los jets originados por los quarks u,d, y s tienen mayor función de fragmentación que los de gluones. Esto quiere decir que sus constituyentes llevan una gran fracción del momento total del jet.[25]

La función de fragmentación puede expresarse de la siguiente forma:

$$p_T D = \frac{\sqrt{\sum_i p_{t,i}^2}}{\sum_i p_{t,i}} \tag{3.4}$$

Así, si toma el valor de 1, quiere decir que sólo una partícula del jet lleva su momento, mientras que si vale cero, el jet estará formado por 'infinitas' partículas que lleven una fracción pequeña de su momento.

En la Figura 3.7 se muestran las distribuciones de los observables multiplicidad, eje menor de la elipse y función de fragmentación para jets originados por gluones y quarks ligeros, tras un

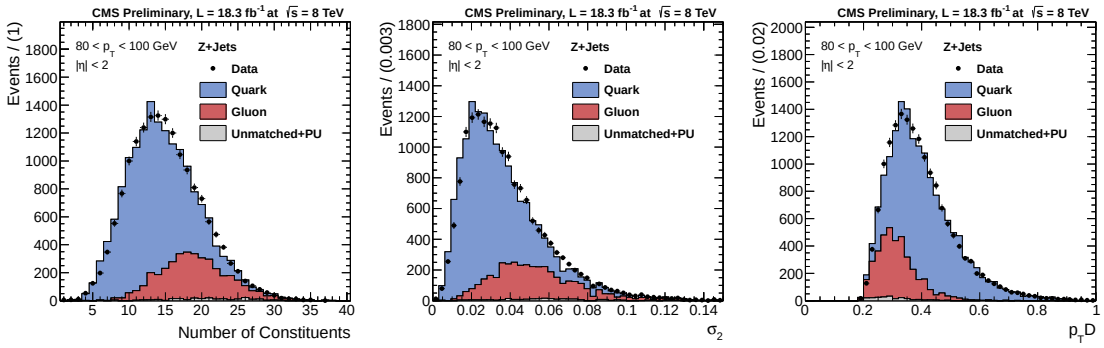


Figura 3.7: Comparación de la distribución de los observables multiplicidad, eje menor de la elipse y función de fragmentación, para jets originados por gluones o quarks ligeros tras un proceso $Z + jets$, con $80 < p_T < 100$ GeV y $|\eta| < 2$. Los puntos negros representan los datos reales. Además, también aparece la distribución para jets no identificados, es decir, del *pile up*. [25]

proceso $Z + jets$ de una colisión protón-protón con energía $\sqrt{s}$=8 TeV. Además, se comparan simulaciones de `MADGRAPH/PYTHIA` frente a los datos reales.

3.2.1. Discriminador de probabilidad

A partir de estos observables físicos, se construye un discriminador de probabilidad, formado por el producto de todos ellos. El uso de este discriminador aporta robustez y simplicidad. Además, se puede interpretar como la probabilidad de que un jet detectado haya sido originado por un quark ligero. Así, el discriminador valdrá uno si el jet detectado viene de un quark ligero y cero si viene de un gluon, aunque al aplicarlo sobre una muestra toma valores contínuos en el intervalo $[0, 1]$. Cabe mencionar que en la construcción del discriminador se utilizó el eje menor de la elipse σ_2, en vez del ancho del jet, por simplicidad en los calculos.[25]

En la Figura 3.8 se muestra la distribución de este discriminador de probabilidad, para jets simulados tras un evento *dijet*, es decir, una colisión que da lugar únicamente a dos jets. Se observa que ambas distribuciones se solapan, lo que indica que el discriminador no es capaz de diferenciar los dos tipos de jets con precisión. Aún así, la distribución para los jets originados por quarks ligeros tiene un máximo en 1, mientras que la distribución para aquellos originados por gluones alcanza el máximo en 0.

El objetivo del trabajo es utilizar una red neuronal profunda para mejorar el resultado de este discriminador, es decir, conseguir unas distribuciones con menor solapamiento.

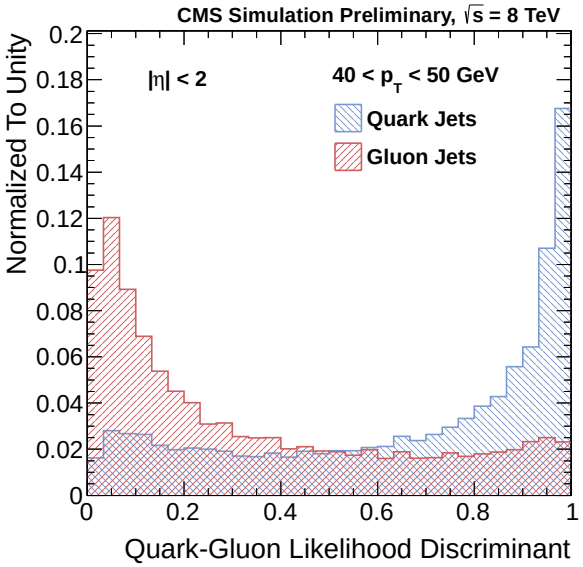


Figura 3.8: Representación gráfica de la distribución del discriminador de probabilidad, para jets originados por quarks ligeros y por gluones, para eventos *dijet* simulados con $40 < p_t < 50$ GeV y $|\eta| < 2$. Se observa que alcanza un máximo cercano a 1 para los jets originados por quarks ligeros, mientras que el valor máximo para los jets generados por gluones es cercano a 0. [25]

Capítulo 4

Aprendizaje automático

Actualmente nos encontramos en la era de los datos. Todas las personas somos generadores diarios de datos y consumidores de los mismos. Como consecuencia, surge la necesidad de analizar estos datos.

Por ejemplo, una cadena de supermercados almacena diariamente los productos comprados por cada cliente. Su objetivo puede ser tener la capacidad de predecir qué productos comprará cada tipo de cliente, de cara a maximizar las ventas y el beneficio, ya que cada cliente quiere encontrar en su supermercado los productos que busca. Sin embargo, encontrar el tipo de personas que compran cierto sabor de helado o un tipo de cereal no es tarea evidente. El comportamiento de los clientes puede cambiar con el tiempo o según la zona geográfica. Sin embargo, sabemos que no es algo completamente aleatorio. Los clientes no acuden al supermercado a comprar cosas aleatorias sino que existen ciertos patrones.

4.1. Introducción al aprendizaje automático

Cuando buscamos resolver un problema a través de un ordenador, utilizamos algoritmos. Un algoritmo es una secuencia de instrucciones que debe ser llevada a cabo para transformar un `input` en un `output`. Sin embargo, puede ocurrir que desconozcamos las instrucciones para transformar las variables de entrada en un `output`. La idea del aprendizaje automático es suplir con datos lo que nos falta por saber. Por ejemplo, si queremos construir un modelo que determine si un correo es *spam* o no, es difícil determinar el algortimo que debemos implementar ya que se trata de una cuestión subjetiva. Sin embargo, se pueden utilizar ejemplos para "*aprender*" qué es un correo *spam* para diferentes personas.

Utilizando aprendizaje automático no identificamos exactamente el proceso, pero tenemos una buena y útil aproximación, a partir de ciertos patrones detectados.

El aprendizaje automático, también conocido como *machine learning* o ML, es una técnica de inteligencia artificial que consiste en asignarle un modelo a una tarea y posteriormente optimizar los parámetros del modelo, utilizando una muestra de datos de entrenamiento o experiencia pasada. Dadas unas variables de entrada $\{x_i\}_{i=1,\dots,n}$ que llamamos `input`, buscamos asignarle una variable de salida o `output` $y \in \mathbb{Y}$ a partir de datos conocidos. Al proceso de utilizar una muestra de datos o experiencia pasada para optimizar los parámetros del modelo se le denomina entrenamiento. Asimismo, a la muestra de datos utilizada en este proceso se le llamará muestra de datos de entrenamiento. Pese a ser considerado una técnica 'oscura', el aprendizaje automático tiene una base estadística y cercana a la inferencia.[26].

Existen tres tipos de aprendizaje automático: aprendizaje supervisado, no supervisado y aprendizaje de refuerzo. En el aprendizaje supervisado, la muestra de entrenamiento son pares `input-output` de la forma $\{x_i, y_i\}_{i=1,\dots,n}$, donde y_i representa un valor del `output` que se quiere predecir de forma general. Por otro lado, el aprendizaje no supervisado se caracteriza por utilizar una muestra de entrenamiento no etiquetada de la forma $\{x_i\}_{i=1,\dots,n}$. Por último, el aprendizaje de refuerzo va más allá, utilizando una muestra de la que no se conoce la etiqueta de salida y realiza el entrenamiento del modelo mediante un sistema de refuerzos y castigos.

El aprendizaje automático resulta de gran utilidad en física de partículas ya que, dada la complejidad teórica de los sucesos que ocurren en el detector, es capaz de ofrecer una aproximación de lo que ocurrirá tras la colisión. En este trabajo se utilizará una muestra generada por simulación de Monte Carlo, donde conocemos el `output` para cada uno de los jets. Por tanto, trabajaremos con aprendizaje supervisado.

4.1.1. Aprendizaje supervisado

La idea básica del aprendizaje automático supervisado consiste en, a partir de una muestra de entrenamiento $\{x_i, y_i\}_{i=1,\dots,n}$, buscar una función $f : \vec{x} \to y$, donde x_i son las variables de entrada y y_i las correspondientes etiquetas. Con el entrenamiento buscamos, a partir de un conjunto de funciones $f_\phi : \mathbb{R}^d \to \mathbb{R}$, optimizar los parámetros ϕ de la función. Cuánto mejor sea dicha aproximación, mayor habilidad tendremos para predecir de forma exacta con datos nuevos. [27]

Dentro del aprendizaje supervisado, diferenciamos entre regresión y clasificación, según el objetivo del modelo y la naturaleza del `output`. Sea $y \in \mathbb{Y}$ la etiqueta que buscamos asignar

a una muestra de datos $\vec{x} \in \mathbb{R}^d$. Si $\mathbb{Y}$ es un espacio discreto, es decir, la variable y sólo puede tomar ciertos valores que llamamos clases, estaremos ante un problema de clasificación. Por otro lado, si $\mathbb{Y}$ es un espacio contínuo, hablaremos de un modelo de regresión. Así, un modelo de clasificación asigna una clase a cada observación muestral.

En nuestro caso, buscamos asignar una clase entre $\{gluon,$ quark u, d y s$\}$ a una muestra de datos simulados, por lo que nos encontramos ante un problema de clasificación. De hecho, en nuestro caso, se tratará de un problema de clasificación binaria. De forma general, se habla de clase positiva y clase negativa.

A continuación se explicarán brevemente diferentes conceptos de los modelos de aprendizaje automático, que facilitan su comprensión.

- Verdadero positivo (TP): Es una observación muestral a la que el modelo le asigna la etiqueta positiva, y acierta.
- Verdadero negativo (TN): Es una observación muestral a la que el modelo le asigna la clase negativa de la variable de salida, y acierta.
- Falso positivo (FP): Es un dato etiquetado como positivo por el modelo, que realmente es negativo.
- Falso negativo (FN): Es un dato clasificado como negativo por el modelo, pero su etiqueta original es positiva.
- Tasa de verdaderos positivos (TPR): representa la fracción de datos positivos etiquetados correctamente.

$$TPR = \frac{TP}{TP + FN} \tag{4.1}$$

- Tasa de falsos positivos (FPR): representa la fracción de negativos mal etiquetados.

$$FPR = \frac{FP}{FP + TN} \tag{4.2}$$

Normalmente, un modelo de aprendizaje automático de clasificación no devuelve una variable binaria que tome valores dicretos $\{0, 1\}$, sino que devuelve una variable y, que toma valores entre $[0, 1]$. Como consecuencia, se establece un umbral $k > 0$, que indica para qué valores de la variable y se pertenece a una clase u a otra. Por ejemplo, los valores superiores o iguales al umbral k se clasificarán como positivos y el resto, como negativos.

A raíz de estos conceptos surge la **curva ROC** *(receiver operating characteristic curve)*, que se utiliza para resumir la relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR), para diferentes valores del umbral k. De esta forma, representa gráficamente una caracterización de la eficiencia del modelo. Al disminuir el umbral, se clasificarán más muestras como positivas, aumentando los verdaderos positivos pero también los falsos positivos.

Para calcular la curva ROC existe un algoritmo llamado AUC *(area under curve)* que calcula el área bajo esta curva. El área bajo la curva ROC proporciona una medida del rendimiento del modelo, para todos los valores posibles del umbral k. Este valor se puede interpretar como la probabilidad de que se clasifique un ejemplo positivo más alto que uno negativo, siendo ambos aleatorios. Si las predicciones de un modelo son todas erróneas, tendrá un valor de AUC de 0, mientras que si acierta en todos los casos, tendrá un AUC de 1. La Figura 4.1 presenta un ejemplo de curva ROC, para una red neuronal profunda.

Por tanto, el valor de AUC mide cuánto de bien se clasifican las predicciones, sin atender a escalas. Sin embargo, si trabajamos con un umbral de clasificación extremo, el AUC puede fallar a la hora de evaluar correctamente el rendimiento del modelo. Un valor del umbral de clasificación extremo se da cuando se quiere penalizar cierta clasificación, es decir, que por ejemplo sea más grave clasificar un positivo como negativo que al revés.

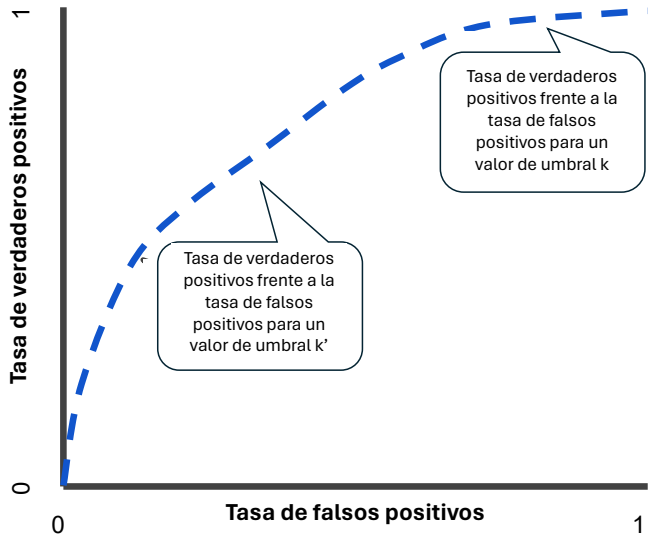


Figura 4.1: Representación gráfica de un ejemplo de curva ROC, para una red neuronal profunda. [28]

Otro concepto importante de los modelos de aprendizaje supervisado es la matriz de confusión. Se trata de una matriz que resume también los conceptos de TP, TN, FP y FN. De forma estándar, tendrá la siguiente forma:

$$\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} \tag{4.3}$$

La matriz de confusión proporciona una medida clara del número de individuos que se clasifican de manera precisa. Los individuos correctamente etiquetados se encuentran en la diagonal principal, mientras que aquellos etiquetados incorrectamente se sitúan en los elementos fuera de esta diagonal. En un modelo perfecto, la matriz de confusión es una matriz diagonal.

Asimismo, resulta importante cuantificar el error que estaríamos cometiendo al implementar el modelo con nuevos datos. Con este objetivo, se definen las siguientes magnitudes:

- **Precisión**: representa la probabilidad de que el modelo acierte con su previsión.

$$P = \frac{TP}{TP + FP} \tag{4.4}$$

  Así, si la precisión vale 0.5 significa que cuando clasifica una muestra como positiva, acierta el 50 % de las veces. Como consecuencia, un modelo que no tiene falsos positivos tiene una precisión de 1.

- **Recuperación o exhaustividad**: representa la fracción de positivos que se identificaron correctamente. Es análogo al ratio de verdaderos positivos.

$$R = \frac{TP}{TP + FN} \tag{4.5}$$

  Así, un modelo que no genera falsos negativos tiene una exhaustividad de 1.

Para evaluar correctamente un modelo, se debe atender tanto a la precisión como a la exhaustividad. Sin embargo, cuando se mejora la precisión de un modelo se suele penalizar la exhaustividad y viceversa. El objetivo será encontrar el mejor equilibrio entre ambas. [29]

Un riesgo o problema del aprendizaje automático es el conocido como sobreentrenamiento o *overfitting*, que ocurre cuando el modelo resultante se ajusta muy bien a la muestra de entrenamiento, pero no predice datos nuevos de forma correcta. Esto se debe a que, al entrenar el modelo, se utiliza una muestra de datos en particular y, por tanto, el modelo resultante se ajustará, lo mejor posible, a estos datos en concreto. Sin embargo, puede ocurrir que al introducir nuevos datos el modelo no generalice bien y no pueda predecir de forma correcta. A menudo, cuando se construye un modelo demasiado complejo, este tiende a generalizar mal y por tanto, resulta poco útil. A menudo, el sobreentrenamiento se produce cuando el modelo "*aprende*" el ruido de la muestra de entrenamiento. En la Figura 4.2 se ilustra el sobreentrenamiento de un modelo frente al bajo entrenamiento.

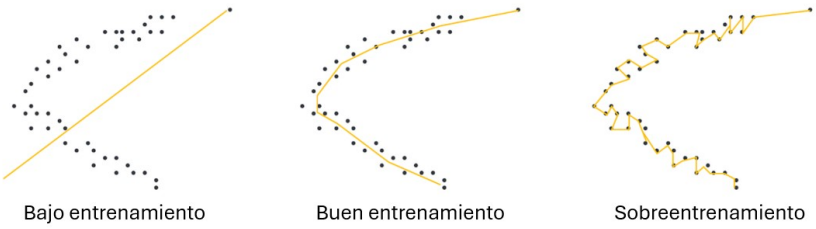


Figura 4.2: Explicación gráfica del sobreentrenamiento de un modelo frente al bajo entrenamiento. [30]

La utilidad de un modelo se mide en su capacidad de generalizar para nuevos datos. Para saber si el modelo construido resulta útil a la hora de predecir nuevos datos, a menudo, se utiliza la validación cruzada. Esta técnica consiste en dividir la muestra de datos de partida en dos, una muestra para el entranamiento y otra muestra que se llamará *test*. Una vez el modelo se ha construido con la muestra de entrenamiento, se evalúa como predice los datos de la muestra *test*. De esta forma, se puede calcular el error que comete el modelo ante nuevos datos, denominado error de validación cruzada. De forma general, el error de validación cruzada será mayor que el error que comete el modelo al predecir la muestra de entrenamiento.

4.2. Redes neuronales profundas

Las redes neuronales son un modelo de aprendizaje automático supervisado basado en el funcionamiento de las neuronas. El cerebro es un potente sistema de análisis de información. Como consecuencia, entender el funcionamiento del cerebro puede ayudar a la creación de algortimos que simulen su funcionamiento en ordenadores. Las neuronas son unidades de procesamiento del cerebro que trabajan en paralelo, con conexiones entre más de 10^4 de ellas. La idea de las redes neuronales es imitar el funcionamiento de las conexiones neuronales del cerebro.

La estructura básica de una red neuronal es una primera capa formada por neuronas de entrada, que reciben información y que transforman en un `output`, que, a su vez, será recogido neuronas en una capa de salida, generando el `output` final de la red. [26]

Las neuronas que constituyen una red neuronal reciben información y la transforman en un único `output`, que toma valores comprendidos entre $[0, 1]$. Además, estas neuronas asignan unos pesos a la información que reciben, que representan su contribución al `output`. En la Figura 4.3 se muestra un ejemplo de una neurona, con variables de entrada x_i. [31]

Cuando entrenamos una red neuronal, esencialmente estamos modificando los pesos asigna-

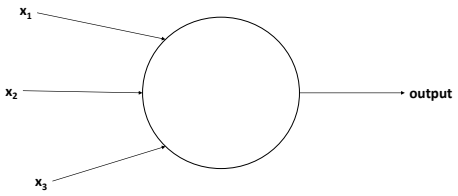


Figura 4.3: Esquema de una neurona de una red neuronal, que recibe 3 variables de entrada x_i y genera un `output`. [31]

dos a cada variable de entrada para cada neurona. Generalmente, una pequeña modificación en los pesos provoca una pequeña modificación en el `output`. Las operaciones que realiza una red neuronal por dentro no son más que funciones matemáticas, y su entrenamiento consiste es aplicar técnicas estadísticas.

Por otro lado, las redes neuronales profundas, o en inglés *Deep Neuronal Network* (DNN), son redes neuronales formadas por más de una capa de neuronas paralelas. Estos modelos son conocidos por ser capaces de aprender distribuciones complejas de muchas dimensiones. La capa inicial de neuronas recibe el nombre de capa de entrada y la última se llamará capa de salida. Asimismo, las capas de neuronas paralelas entre la capa de entrada y de salida se denominan capas internas u ocultas. Cuantas más capas internas tenga la red neuronal, más compleja será su estructura. Las neuronas de las capas internas reciben la información del `output` de las neuronas de la capa anterior. El número de neuronas en cada capa junto con el número de capas determina la estructura de nuestro modelo.

En la Figura 4.4 se muestra un ejemplo de una red neuronal profunda, con dos capas internas. En este caso aparece una única neurona en la capa final, pero podría haber tantas como se quiera.

Si el objetivo de la red neuronal es realizar una clasificación binaria, generalmente, se utiliza una única neurona de salida, que devolverá un valor comprendido entre 0 y 1. Así, si el `output` es mayor que k, el umbral previamente establecido, significará que esa muestra pertenece a la clase 1 de la variable que se busca predecir, mientras que si es menor, será de la clase 0.

En nuestro caso, buscamos realizar una clasificación binaria entre si el jet fue originado por un quark ligero o un gluon. Como consecuencia, se trabajará con redes neuronales profundas con una única neurona en la capa de salida.

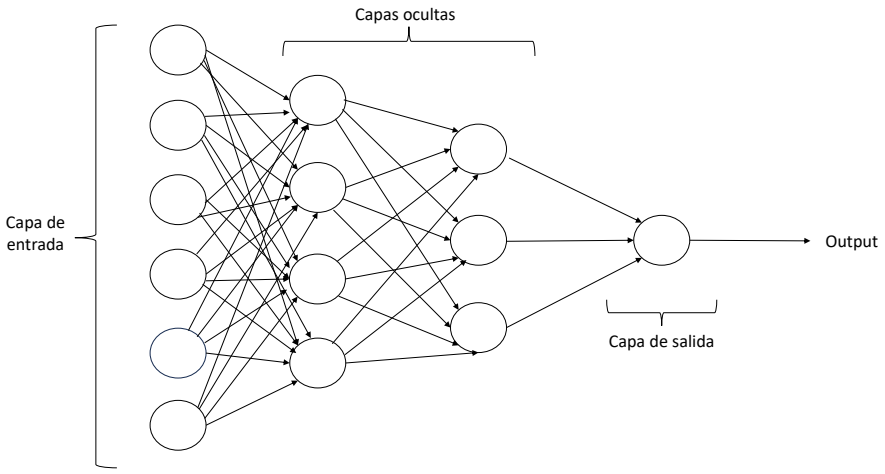


Figura 4.4: Esquema de un ejemplo de una red neuronal profunda, formada por dos capas internas u ocultas. [31]

4.2.1. Dropout

Como se ha comentado previamente en este capítulo, un riesgo de los modelos de aprendizaje automático es el sobreentrenamiento. En el caso de un modelo de redes neuronales profundas, existe una técnica para reducir este sobreentrenamiento, llamada *dropout*.

El sobreentrenamiento de una red neuronal se produce cuando esta aprende el ruido de la muestra de entrenamiento y, como consecuencia, generaliza mal a la hora de predecir nuevos datos. En la Figura 4.2 se muesta un ejemplo gráfico del sobreentrenamiento de un modelo.

La mejor forma para evitar el sobreentrenamiento es utilizar una muestra de entrenamiento extensa que abarque todos los posibles valores de cada variable de entrada, pero esto resulta muy costoso computacionalmente.

La idea del *dropout* es la eliminación de algunas neuronas de la red con cierta probabilidad p. Cuando se entrena una red neuronal, se busca optimizar la predicción de la muestra de entrenamiento. Sin embargo, puede ocurrir que una neurona "*aprenda*" demasiado e incorpore el ruido estadístico provocando el sobreentrenaminto de la red. Al utilizar el *dropout*, las neuronas se tienen que adaptar a la pérdida de una de ellas y por tanto, se reduce la probabilidad de que una de ellas aprenda demasiada información. El funcionamiento del *dropout*, consiste en ir desconectando neuronas durante el entrenamiento de la red hasta encontrar la combinación óptima de las mismas. En la Figura 4.5, se presenta un ejemplo gráfico del resultado de aplicar *dropout* en una red neuronal.[32]

Si por ejemplo, establecemos un *dropout* de 0.2 para cada capa de neuronas, esto quiere

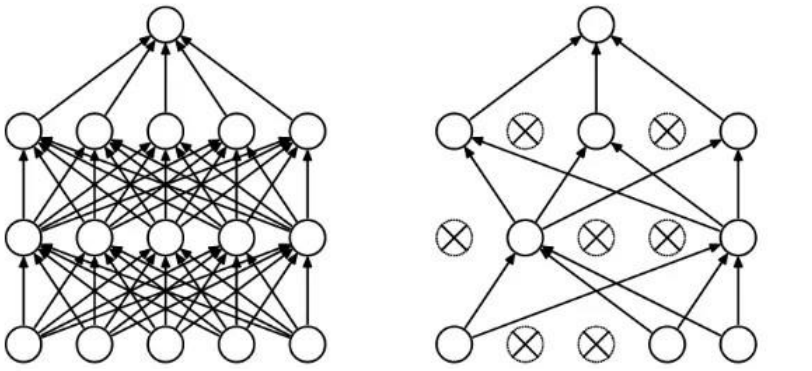


Figura 4.5: Ejemplo gráfico del efecto de aplicar la técnica del *dropout* en una red neuronal. [32]

decir que durante el entrenamiento se desconectarán el 20 % de las neuronas de cada capa, hasta encontrar la opción óptima.

Como consecuencia, al utilizar la técnica del *dropout* en una red neuronal, se obtendrá un modelo más simple, que prediga peor la muestra de entrenamiento pero que generalice mejor ante nuevos datos.

Capítulo 5

Muestra de datos

En este capítulo se profundizará sobre la simulación y obtención de la muestra. Además, se explicará la estructura de los datos con los que se trabajará, y se mostrarán las distribuciones de los observables más relevantes, para una mejor comprensión de los resultados.

5.1. Obtención de los datos

Para la construcción del modelo de aprendizaje automático, se utilizará una muestra de datos que representa una simulación de los choques protón-protón producidos en el experimento CMS, con una energía de centro de masas de 13 TeV. La muestra ha sido generada mediante el método de Monte Carlo y, posteriormente, se le ha aplicado un algoritmo para simular el detector CMS. Cabe mencionar que se utiliza una base de datos simulada porque el modelo necesita conocer el origen de cada jet para su entrenamiento. [33]

En primer lugar, los datos son generados utilizando un generador de sucesos de Monte Carlo llamado *Pythia 8*. [19]. Este software se utiliza, entre otras cosas, para la descripción de colisiones de partículas a altas energías. Está basado en modelos teóricos como la interacción débil y fuerte, las distribuciones de partones o las cascadas de partones y utiliza un algoritmo numérico para producir secuencias aleatorias de eventos. Como los choques de partículas reales son eventos estocásticos, para poder simularlos, se necesita una generación de números pseudoaleatorios. De forma resumida, cada número pseudoaleatorio generado se compara con una función de densidad acumulada, de manera que determina un ángulo de la partícula, el tipo de partícula resultante de la hadronización, etc. Este tipo de técnicas resultan de gran utilidad para estudiar estos fenómenos y saber si, realmente, se están identificando correctamente las partículas.

Así, el software *Pythia 8* nos proporciona unos datos que simulan el resultado de una colisión

protón-protón ocurrida en un colisionador, es decir, obtenemos un conjunto de jets de QCD. A continuación, se aplica el simulador *GEANT4*, que reproduce las señales eléctricas que estos jets dejarían en el detector CMS. [34] Posteriormente, el algoritmo *Particle-Flow* devuelve las partículas reconstruidas a partir de las señales del detector. Este algoritmo utiliza la información de todos los subdetectores para reconstruir una lista de estados finales y su trayectoria, que reciben el nombre de *PFCandidates*. Además, también es capaz de mitigar el ruido. Su funcionamiento se puede resumir en conectar los depósitos de energía en diferentes subdetectores para obtener la traza de una partícula y, posteriormente, utilizar estos datos para deducir su tipo y energía.[35]

Por último, se aplica el algoritmo $Anti - K_t$ a las *PFCandidates* obtenidas. Este se encarga de agrupar las partículas detectadas para formar jets. Al igual que el algoritmo *Particle-Flow*, se utiliza tanto para los datos simulados como para los reales. Normalmente, se toma el parámetro de radio R=0.4, que representa el radio del cono de las partículas que componen un jet, en el plano $\eta - \phi$. Por tanto, el valor de R indica cómo de juntas están las trazas que se considera que pertenecen al mismo jet. En particular, la muestra utilizada en este trabajo fue tratada con un valor de R=0.4. Para formar los jets, el algoritmo $Anti - K_t$ prioriza la agrupación de una partícula pesada con las partículas ligeras que se encuentren en su proximidad. Así, si por ejemplo una partícula pesada no encuentra a otra en una distancia $2R$, acumulará todas las partículas ligeras en un radio R, formando un jet con una forma de cono perfecta. La Figura 5.1 muestra un ejemplo gráfico de la agrupación de partículas en jets, mediante el algoritmo $Anti - K_t$. [36]

Más aún, se han aplicado las correcciones L1, L2 y L3 para mitigar la energía residual de los jets. La correción L1 consiste en restar la contribución media del momento transverso correspondiente al *pile up* en el área del cono del jet. Esta contribución media varía en función de la pseudorapidez. Por otro lado, las correciones L2 y L3 ajustan el momento medido del jet al real, utilizando técnicas de equilibrio de momento y energía perdida, en eventos *dijet*. Posteriormente se ha reducido el *pile up* utilizando el algoritmo CHS, *charged hadron subtraction*, que consiste en eliminar los hadrones cargados cuya traza reconstruida se origina en vértices asociados con el *pile up*. [37] [38]

Además, sólo se consideran los jets con $p_t > 30$ GeV y $|\eta| < 2.5$, dado que los jets de mayor interés serán los más energéticos y estos suelen producirse a valores bajos de η. Para determinar el sabor de cada jet, se utiliza el *flavor algorithm*, que intenta asociar cada jet con una única partícula inicial, para después asignarle el sabor de dicha partícula. En concreto, se utilizará la definición *physics*, que busca gluones y quarks resultantes de la colisión inicial a partir de la

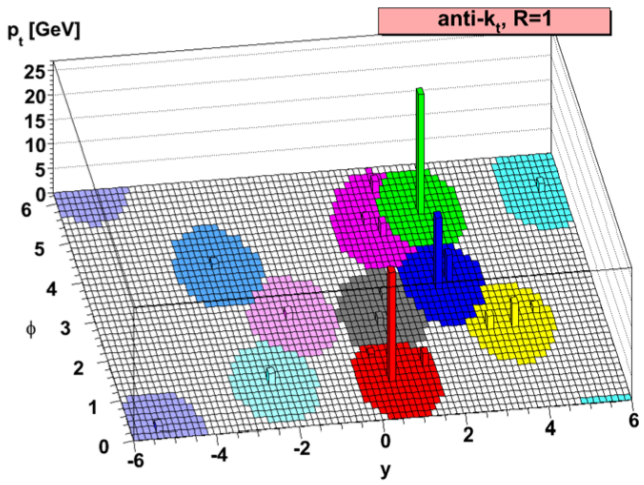


Figura 5.1: Esquema de la agrupación de partículas en jets, utilizando el algortimo *anti-*k_t, con R=1, para una muestra generada con Herwig[8]. [36]

información del generador *Pythia 8*.

Como nota, si se trataran de datos reales, se partiría de las señales de los detectores y se les aplicaría los algoritmos *Particle Flow* y $Anti-K_t$ para obtener los jets resultantes.

5.2. Análisis descriptivo de la muestra

A continuación, se describirá la muestra a tratar, desde su estructura a la distribución de los observables más importantes, con el fin de entender mejor los datos para su posterior análisis.

La muestra consta de información a tres niveles diferentes: nivel alto, nivel de partícula y nivel de generador. El nivel alto o RECO contiene la información obtenida tras aplicar el algortimo $Anti-k_t$, es decir, la información reconstruida que buscaríamos analizar en el experimento CMS. Por otro lado, el nivel de partícula muestra la información de los *PFcandidates*, obtenida tras aplicar el algortimo *Particle-Flow*, a la información que simula las señales del detector CMS. Por tanto, se trata de la información que obtendríamos del experimento. Por último, el nivel de generador es lo que buscamos predecir. En este caso, es previamente conocido, ya que se trata de una muestra simulada. Sin embargo, en el experimento desconoceríamos la información a nivel de generador. Como consecuencia, esta información no podrá ser usada en el entrenamiento de la red neuronal. En la Figura 5.2, se presenta un esquema de la información recopilada en la muestra y sus distintos niveles.

Por otro lado, la matriz de datos cuenta con una columna para cada observable y una fila

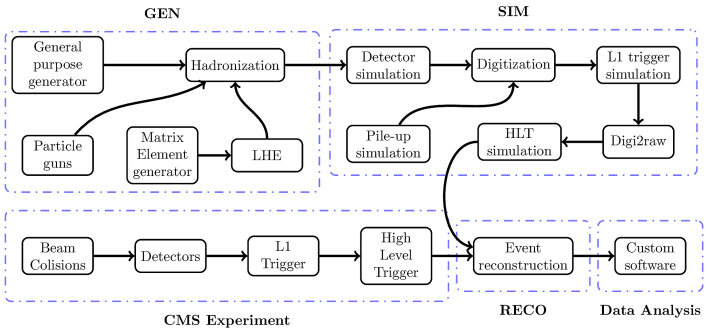


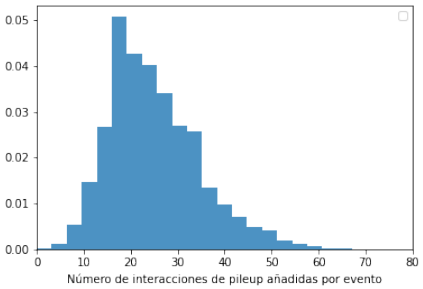
Figura 5.2: Esquema de la información recogida en la muestra simulada. En primer lugar, está la información de nivel de generador (GEN), obtenida del simulador de eventos, en nuestro caso *Pythia 8*. A continuación, está el nivel de partícula o simulación (SIM), que se corresponde con la información que obtendríamos del experimento. Por último, están los datos reconstruidos o nivel alto (RECO), que serán analizados. [39]

para cada jet simulado. Además, el número de eventos en nuestra muestra es de 2.2554294 10^7, donde cada evento es una colisión protón-protón principal, es decir, de una gran energía. Sin embargo, dada la memoria del dispositivo con el que se ha trabajado, sólo se han utilizado 17GB de información de los 190GB disponibles.

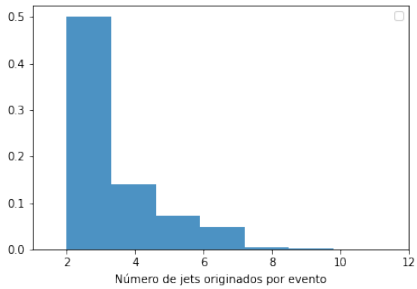
Ahora bien, los generadores de sucesos simulan únicamente la colisión principal o *hard scattering* del proceso de interés. Por tanto, la muestra simulada no contiene información del *pile up*. Como consecuencia, el *pile up* es añadido a la muestra, para ser analizado también. En la Figura 5.3a se muestra la distribución del número de interacciones del *pile up* agregadas, para un evento de la muestra simulada con la que se está trabajando. El número de interacciones del *pile up* depende de la luminosidad instantánea del detector, es decir, de las condiciones del LHC en el momento de la toma de datos. En la Figura 2.2 se muestra la evolución del número de interacciones del *pile up* con los años, al aumentar la luminosidad del LHC. La muestra simulada con la que se trabaja, representa una toma de datos del Run 2 del LHC, que tuvo lugar entre 2015 y 2018. Por tanto, el número de interacciones del *pile up* agregadas debería ser similar al apilamiento detectado en esta toma de datos.

Asimismo, cada evento suele dar lugar de media a dos o tres jets. En particular, en la muestra utilizada, la media es de 3.3 jets de más de 30 GeV y $|\eta|$ <2.5. En la Figura 5.3b se muestra la distribución del número de jets resultantes, para los eventos estudiados.

A la hora de construir la red neuronal, de entrada, buscamos utilizar aquellos observables cuya distribución para los dos tipos de jets se solapen lo menos posible. Como se ha comentado



(a) Distribución del número de interacciones del *pile up* que son añadidas a un evento de la muestra.



(b) Distribución del número de jets resultantes de un evento para la muestra estudiada.

Figura 5.3: Distribuciones del número de interacciones del *pile up* añadidas y el número de jets resultantes en los eventos en la muestra estudiada.

en el capítulo 3, los jets originados por gluones o por quarks ligeros, muestran diferencias en sus respectivas distribuciones de multiplicidad, eje menor de la elipse y función de fragmentación. De hecho, estos observables constituyen el discriminador de probabilidad, cuyos resultados se compararán, con los obtenidos mediante la red neuronal. La distribución de estos observables, para la muestra simulada, se observa en la Figura 5.4.

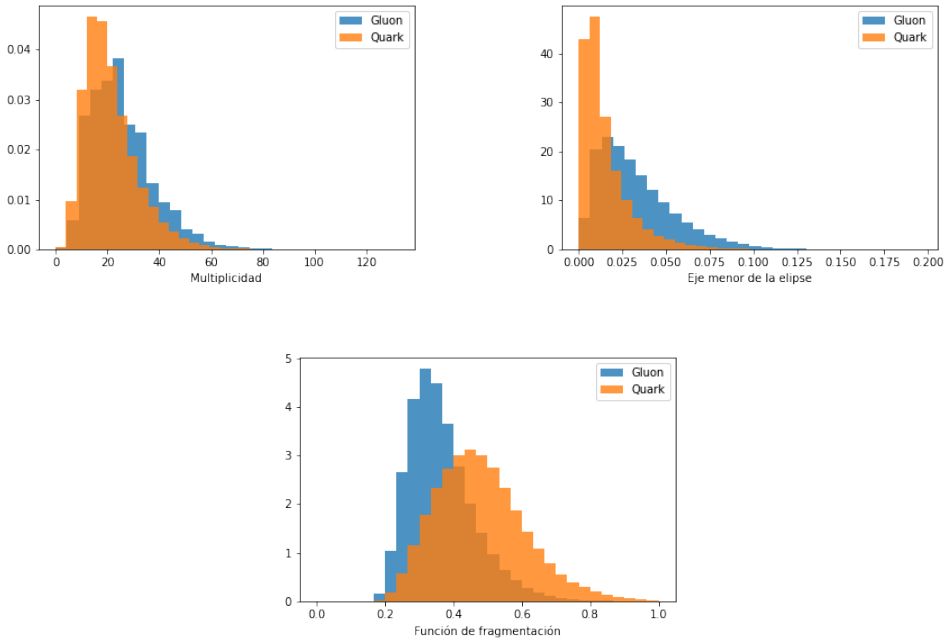


Figura 5.4: Comparación de la distribución de los observables multiplicidad, eje menor de la elipse y función de fragmentación para la muestra de jets, simulados por el método de Monte Carlo. Las tres son magnitudes adimensionales.

Después de analizar las representaciones gráficas de la Figura 5.4, se verifica que los jets originados por un quark ligero presentan una menor multiplicidad. Además, se observa que el eje menor de la elipse es mayor para aquellos jets originados por gluones. También se confirma que los jets de quarks *u, d y s* tienen mayor función de fragmentación. Estos resultados respaldan la teoría expuesta en el capítulo 3. Adicionalmente, la distribución del discriminador de probabilidad para la muestra simulada, se muestra en la Figura 5.5. Comparando con la Figura 3.7, se observa una notable similitud.

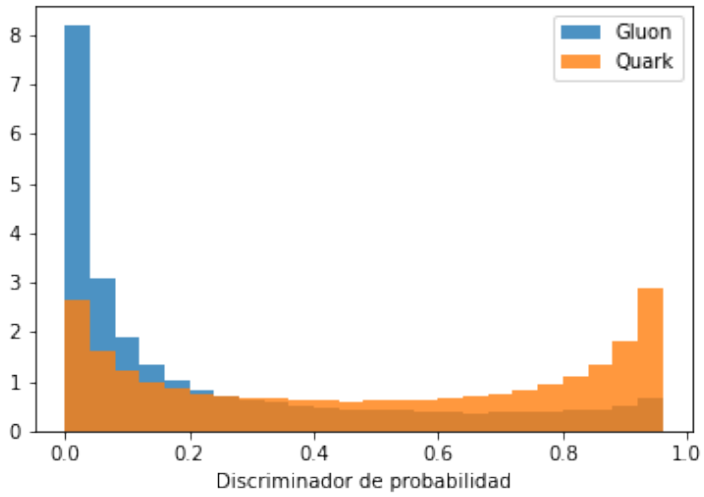


Figura 5.5: Distribución del discriminador de probabilidad, definido en el capítulo 3, para la muestra estudiada. Se trata de una magnitud adimensional por definición.

Tras la representación de diferentes observables de la muestra a nivel alto, se observó que había tres de ellos que destacaban, por tener el menor solapamiento entre ambas distribuciones. Estos observables son la función de fragmentación, el eje menor de la elipse y el *girth* o momento lineal radial, que se define según la ecuacion 5.1.

$$g = \sum_{i \in jet} \frac{p_t^i}{p_t^{jet}} |r_i| \tag{5.1}$$

donde $|r_i| = \sqrt{\Delta y_i^2 + \Delta \phi_i^2}$. Igualmente, se trata de una magnitud adimensional. [40]

En la Figura 5.6, se presenta la distribución del *girth* para jets de gluones y quarks ligeros, en la muestra simulada.

La importancia del eje menor de la elipse y de la función de fragmentación era esperada, ya que estos aspectos están respaldados por la teoría discutida en el capítulo 3. Adicionalmente, el *girth* podría considerarse un observable significativo en la discriminación de los jets.

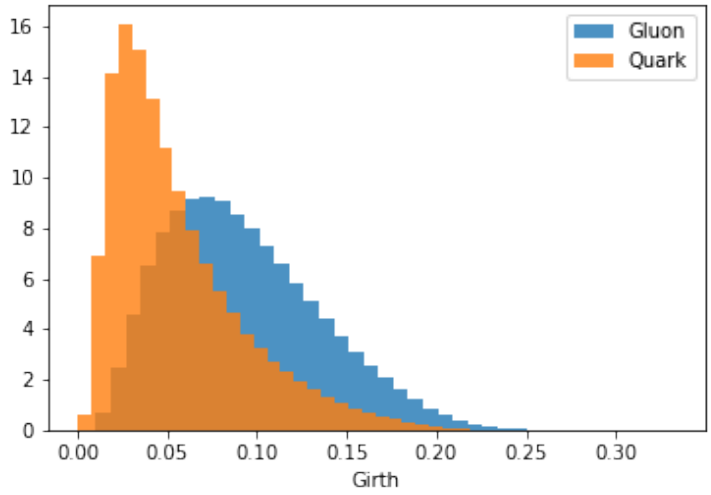


Figura 5.6: Distribución del *girth* de jets originados por gluones o quarks ligeros de la muestra simulada.

Finalmente, la Figura 5.7 presenta las distribuciones de otros observables de la muestra de datos. Por una parte, definimos el área de un jet como la superficie de la elipse que forma su proyección en el plano $\eta - \phi$. Además, aparecen la pseudorapidez η del jet y su multiplicidad de hadrones cargados y neutros, es decir, el número de hadrones cargados y neutros respectivamente en el jet. La masa del jet será la masa total del mismo, una vez aplicadas ciertas correcciones de energía, y se mide en GeV. Por último, el momento transverso del jet es la suma total de todas sus componentes, y se mide en GeV. Además, en la Figura 5.7, se evidencia que, observables como la multiplicidad de hadrones cargados o el momento transverso, tienen distribuciones diferentes para cada tipo de jet. Por el contrario, la distribución de la masa no presenta diferencias significativas entre los jets originados por gluones y quarks ligeros. Como consecuencia, en un primer instante, parece que los observables multiplicidad de hadrones cargados y momento transverso del jet, tendrán un papel más significativo en la clasificación de los jets.

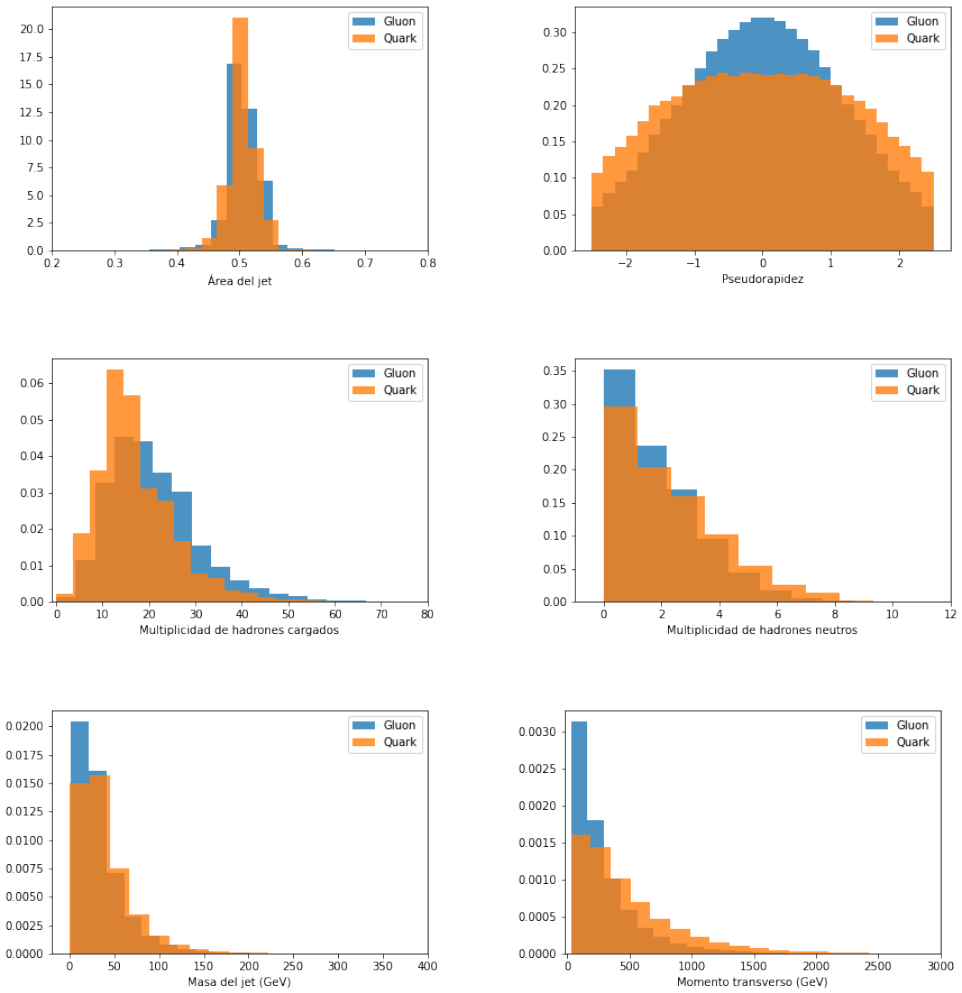


Figura 5.7: Comparación de la distribución del área del jet, la pseudorapidez, la multiplicidad de hadrones cargados en el jet, la multiplicidad de hadrones neutros, la masa del jet y el momento transverso del jet, para la muestra de jets, simulados por el método de Monte Carlo.

Capítulo 6

Resultados

A continuación, se expondrán los resultados obtenidos tras la construcción de una red neuronal profunda, para el análisis de la muestra simulada de Monte Carlo.

Inicialmente, se partirá de una red neuronal profunda, formada por 100 neuronas en la capa de entrada, dos capas ocultas con 100 y 50 neuronas respectivamente y, finalmente, una única neurona en la capa de salida. Además, la red cuenta con dos `dropout` de probabilidad 0.2, previos a cada una de las capas ocultas. Asimismo, se ha configurado el entrenamiento de la red con 20 épocas o etapas y, al inicio de cada etapa, se barajan los datos, para que la red no memorice su orden. Por ende, la red obtendrá diferentes resultados en cada entrenamiento.

La red neuronal buscará predecir si cada jet de la muestra ha sido originado por un quark ligero o un gluon. Para ello, se define una variable binaria, que tomará el valor 1 cuando el jet haya sido originado por un quark ligero, y 0 en el caso de un gluon. La red devuelve el valor de esta variable para cada jet de la muestra. Considerando un umbral de 0.5, los jets que obtengan un valor mayor, serán clasificados como originados por quarks ligeros. De ocurrir lo contrario, el modelo asignará la etiqueta de gluón, al origen del jet. Este mismo umbral se utiliza para construir la matriz de confusión de la red, que cuantifica el número de jets de cada clase correctamente etiquetados.

Debido a limitaciones computacionales, se ha utilizado únicamente el 9 % de la muestra simulada disponible. Además, los datos utilizados se dividieron en una muestra de entrenamiento, que supuso el 85 %, y muestra test, formada por el 15 % restante. También resaltar que, previamente, la muestra ha sido filtrada, eliminando todos aquellos jets que no hayan sido originados por un quark ligero o un gluon.

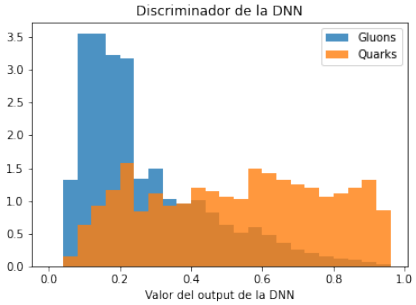
En primer lugar, se han comparado los resultados obtenidos con la red neuronal frente al discriminador de probabilidad. Seguidamente, se ha estudiado la variación de los resultados en

función de las variables de entrada y la arquitectura de la red. Por último, se ha estudiado si al invertir la variable de salida de la red, se obtienen los mismos resultados. En otras palabras, se ha utilizado una variable binaria que valga 1 cuando el jet haya sido originado por un gluon, y 0 en caso de ser generado por un quark.

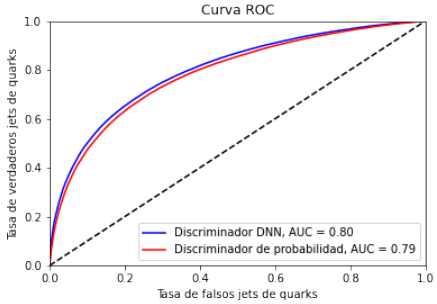
6.1. DNN frente al discriminador de probabilidad

Primeramente, se buscará determinar si el uso de una red neuronal profunda para clasificar jets, es capaz de mejorar el resultado obtenido con el discriminador de probabilidad descrito en el capítulo 3. Para ello, se empezará trabajando únicamente con los observables utilizados en la construcción de dicho discriminador, es decir, el eje menor de la elipse, la multiplicidad y la función de fragmentación.

En la Figura 6.1 se muestran los resultados obtenidos con el modelo de aprendizaje automático. Por una parte, se observa que la distribución del discriminador de probabilidad no alcanza un máximo en 1 para los jets originados por quarks ligeros, siendo este el valor esperado para esta clase de jets. En contraste, en el caso de los gluones, se alcanza un máximo cercano a 0, como se esperaba. Aún así, el área bajo la curva ROC del discriminador obtenido con la DNN es 0.01 mayor, lo que indica una ligera mejoría con respecto al discriminador de probabilidad.



(a) Distribución del discriminador construido con la red neuronal profunda. En la Figura 5.5 se muestra el discriminador de probabilidad para la misma muestra.



(b) Representación gráfica de la curva ROC para la red neuronal profunda y para el discriminador de probabilidad. Se obtienen respectivamente unos valores de AUC de 0.80 y 0.79.

Figura 6.1: Resultados obtenidos con la red neuronal, utilizando como variables de entrada el eje menor de la elipse, la multiplicidad y la función de fragmentación, frente al discriminador de probabilidad.

Por otro lado, la matriz de confusión obtenida fue la siguiente:

$$\begin{pmatrix} 0.54 & 0.08 \\ 0.17 & 0.21 \end{pmatrix} \tag{6.1}$$

donde la primera fila hace referencia a los gluones, mientras que la segunda se refiere a los quarks ligeros. Además, la matriz de confusión se presenta en proporciones, para facilitar la lectura. En este contexto, la primera entrada indica que el 54 % de los jets son generados por gluones y están etiquetados de manera correcta. De la misma manera, se observa que el 21 % de los jets son originados por quarks ligeros y están clasificados de manera precisa.

Por consiguiente, utilizando la red neuronal, se comete un error de validación cruzada del 25 %, es decir, al utilizar la red neuronal como discriminador, el 25 % de la muestra test se etiqueta erróneamente. En otras palabras, la precisión del modelo es 0.75. Sin embargo, utilizando el discriminador de probabilidad se comete un error del 27 %. Como consecuencia, se han mejorado los resultados.

Además, con el modelo de la red neuronal profunda, se obtiene una precisión para los jets generados por gluones de 0.76 frente a un valor de 0.72 para los quarks, lo que significa que es más probable que la red clasifique correctamente un jet originado por un gluon.

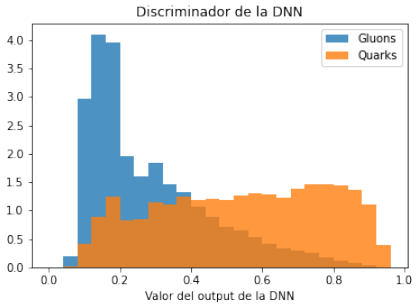
Asimismo, el modelo tiene una exhaustividad de 0.87 y 0.55 respectivamente, que implica que la fracción de jets de quarks ligeros etiquetados correctamente, es menor que la de los gluones. Por tanto, destaca que el modelo es capaz de identificar mejor los jets originados por gluones, frente a aquellos generados por quarks ligeros.

6.2. Variables de entrada

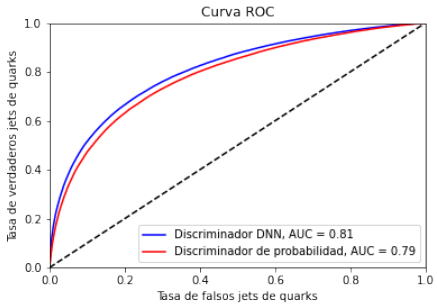
La red neuronal profunda permite utilizar todos los observables de la muestra, sin limitarse a aquellos utilizados en el discriminador de probabilidad. A continuación, se estudiará la variación de los resultados, al modificar los observables de entrada del modelo. El objetivo será encontrar el conjunto de variables que optimice el entrenamiento del modelo.

Como se mostró en el capítulo 5, algunos observables tienen una distribución diferente, o con menor solapamiento, para jets originados por gluones y quakrs ligeros. Con el fin de mejorar el rendimiento de la red neuronal, se añadirán estas variables. Sin embargo, utilizar estas variables no tiene por qué mejorar el rendimiento. Puede ocurrir que existan ciertas conexiones entre observables, a simple vista no identificables, que la red neuronal detecte y que funcionen mejor en la clasificación.

En el capítulo anterior, se comprobó que los observables cuyas distribuciones se solapan menos para cada clase de jets, eran el *girth*, la función de fragmentación y el eje menor de la elipse. Como consecuencia, parece lógico utilizar estos observables para entrenar la red neuronal. Utilizando las variables del discriminador de probabilidad junto con el *girth* se obtuvieron los resultados presentados en la Figura 6.2.



(a) Distribución del discriminador construido con la red neuronal profunda.



(b) Representación gráfica de la curva ROC para la red neuronal profunda y para el discriminador de probabilidad. Se obtienen respectivamente unos valores de AUC de 0.81 y 0.79.

Figura 6.2: Comparación de los resultados obtenidos con los observables *girth*, eje menor de la elipse, función de fragmentación y multiplicidad.

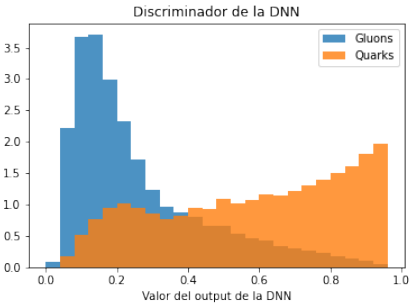
Además, la matriz de confusión será 6.2, donde la primera fila hace referencia a los jets originados por gluones. A partir de esta, se obtiene una precisión de 0.76 y 0.75, para jets originados por gluones y quarks ligeros respectivamente. Por otro lado, se obtuvo una exhaustividad de 0.89 y 0.53 respectivamente. Así, el 89 % de los jets generados por gluones son etiquetados correctamente, frente al 53 % de los jets de quarks ligeros. Por tanto, la red sigue etiquetando mejor los jets originados por gluones.

$$\begin{pmatrix} 0.56 & 0.07 \\ 0.17 & 0.20 \end{pmatrix} \tag{6.2}$$

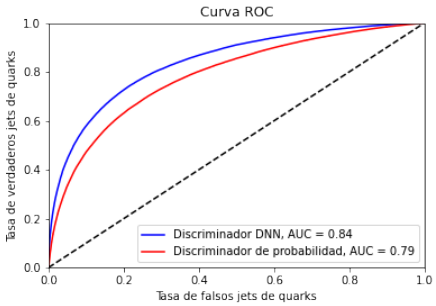
Asimismo, el error de validación cruzada cometido por la red neuronal es del 24 %. Como consecuencia, al añadir simplemente la variable *girth* en el entrenamiento de la red, hemos mejorado el resultado, alcanzando una precisión de 0.76.

Ahora bien, la muestra simulada contiene 21 variables que podríamos utilizar para entrenar la red neuronal, dado que son observables que conoceríamos si se tratara de una muestra real. Por tanto, existen muchísimas agrupaciones de las mismas. A partir de las distribuciones representadas en la Figura 5.7, se utilizaron las siguientes variables en el entrenamiento de la red: η,

función de fragmentación, eje menor de la elipse, *girth*, área del jet, multiplicidad del jet, masa del jet, multiplicidad de hadrones cargados y multiplicidad de hadrones neutros. En la Figura 6.3 aparecen representadas las gráficas del discriminador y la curva ROC obtenidas.



(a) Distribución del discriminador construido con la red neuronal profunda.



(b) Representación gráfica de la curva ROC para la red neuronal profunda y para el discriminador de probabilidad. Se obtienen respectivamente unos valores de AUC de 0.84 y 0.79.

Figura 6.3: Resultados obtenidos con las variables de entrada η, función de fragmentación, eje menor de la elipse, *girth*, área del jet, multiplicidad del jet, masa del jet, multiplicidad de hadrones cargados y multiplicidad de hadrones neutros.

En este caso, se evidencia que la distribución del discriminador alcanza un máximo cercano a 1 para los jets generados por quarks y en torno a 0 para los jets originados por gluones, lo cual concuerda con los valores esperados en cada caso. Además, el área bajo la curva ROC ha alcanzado un valor de 0.84, superando en 0.05 al discriminador de probabilidad.

Asimismo, la matriz de confusión obtenida fue:

$$\begin{pmatrix} 0.54 & 0.09 \\ 0.13 & 0.24 \end{pmatrix} \tag{6.3}$$

Al igual que antes, la matriz de confusión se expresa en proporciones.

En este caso, se registró un error de validación cruzada de 22 %. Además, la precisión fue de 0.80 para los gluones y de 0.74 para los quarks ligeros, mientras que la exhaustividad alcanzó los valores 0.87 y 0.63 respectivamente. Así pues, la red neuronal sigue clasificando de forma más eficaz los jets originados por gluones.

Análogamente, existen otras agrupaciones de variables con las que se consigue también un AUC de 0.84, puesto que existen inifinitas curvas con un AUC de 0.84 que, para un valor del umbral de k=0.5, pasen por el mismo punto que la cruva ROC obtenida en la Figura 6.3.

En conclusión, el rendimiento de la red neuronal depende de los observables de entrada. Como consecuencia, escogiendo unos observables con el suficiente poder discriminatorio entre ambos jets, se mejoran los resutados considerablemente. No obstante, no existe una única combinación óptima de los mismos, sino que se han encontrado varias agrupaciones de observables que dan lugar al mismo resultado.

Además, el modelo permite conocer qué variables utilizan más las neuronas para generar su `output`, lo que puede arrojar luz sobre las mejores variables de entrada a utilizar. En concreto, la red neuronal asocia a cada observable un peso para cada capa de la red. Este peso se calcula como la media de la contribución de la variable al `output` de cada neurona de una determinada capa, y se puede interpretar como su importancia. Sin embargo, esta magnitud no debe interpretarse como algo lineal o unilateral, ya que una variable con muy poco peso asociado puede ser indispensable para la clasificación.

A continuación se presentan los pesos asociados al último conjunto de variables, con las que se alcanza el mejor resultado.

Variable	Peso en la 1 capa oculta	Peso en la 2 capa oculta
Masa	0.54	0.22
η	0.52	0.14
Función de fragmentación	0.50	0.32
Multiplicidad	0.49	0.34
Girth	0.49	0.20
Multiplicidad de hadrones cargados	0.45	0.31
Área	0.45	0.28
Eje menor de la elipse	0.44	0.24
Multiplicidad de hadrones neutros	0.37	0.21

Cuadro 6.1: Tabla con los pesos asociados a cada variable para las capas ocultas. Las variables aparecen ordenadas de mayor a menor peso en la primera capa oculta.

En la tabla 6.1 se observa que las variables más importantes, de acuerdo con los pesos, fueron la masa del jet y la multiplicidad, para la primera capa y la segunda respectivamente. Es oportuno mencionar que no se detallan los pesos de la capa de entrada ya que pueden estar sujetos a la ordenación aleatoria de las variables, al comienzo de cada etapa.

6.3. Estructura de la red neuronal

A continuación, se estudiará los efectos que produce la alteración de la estructura de la red neuronal sobre su rendimiento. Se llevarán a cabo modificaciones de manera individual en cada una de las características, con el fin de realizar una comparación que refleje de manera realista los cambios. Como variables de entrada del modelo, se utilizarán η, función de fragmentación, eje menor de la elipse, *girth*, área del jet, multiplicidad del jet, masa del jet, multiplicidad de hadrones cargados y multiplicidad de hadrones neutros, ya que han demostrado obtener los mejores resultados.

En primer lugar, se notó que tanto la pérdida como la precisión del modelo, durante el entrenamiento, se estabilizaban aproximadamente en la décima época. Esto sugiere la posibilidad de sobreajuste del modelo. En consecuencia, al reducir el número de épocas, se podrían obtener resultados equivalentes, en un periodo de tiempo menor. Al ejecutar el código con tan solo 10 épocas, se confirmó que los resultados se mantienen inalterados. Como consecuencia, en los posteriores resutlados se utilizarán sólamente 10 etapas en el entrenamiento de la red.

6.3.1. Número de capas ocultas

Inicialmente, se estudiará la variación de loss resultados en función del número de capas ocultas del modelo. En este trabajo, se empezó utilizando dos capas ocultas, de 100 y 50 neuronas respectivamente. A continuación, se procederá a estudiar si se trata de una combinación óptima.

En primer lugar, se modificará el número de capas ocultas con 100 neuronas, manteniendo la última capa de 50 neuronas. En la tabla 6.2 se muestran los resultados.

Capas ocultas con 100 neuronas	AUC	Error	Precisión Gluones	Precisión quarks
0	0.84	22 %	0.80	0.75
1	0.84	22 %	0.80	0.75
2	0.84	22 %	0.80	0.75
3	0.84	22 %	0.79	0.76

Cuadro 6.2: Variación del AUC, el error de validación cruzada y la precisión de la red en función del número de capas ocultas con 100 neuronas.

Se observa que no se producen alteraciones al añadir capas ocultas con 100 neuronas. Sin embargo, cuantas más capas de neuronas, mayor tiempo de entrenamiento requiere la red. Como

consecuencia, parece que la estructura óptima es tener una o ninguna capa oculta con 100 neuronas.

Análogamente, se examinó la influencia del número de capas ocultas de 50 neuronas sobre el rendimiento de la red. En este caso, se mantuvo la capa oculta inicial de 100 neuronas. En la tabla 6.3 se exponen los resultados obtenidos.

Capas ocultas con 50 neuronas	AUC	Error	Precisión Gluones	Precisión quarks
0	0.84	22 %	0.79	0.76
1	0.84	22 %	0.80	0.75
2	0.84	22 %	0.81	0.74
3	0.84	22 %	0.81	0.73

Cuadro 6.3: Variación del AUC, el error de validación cruzada y la precisión de la red, en función del número de capas ocultas con 50 neuronas.

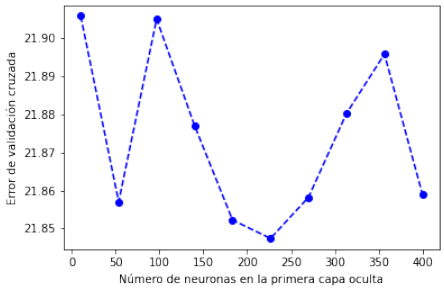
Igualmente, añadir capas de 50 neuronas no mejora los resultados de la red, mientras que aumenta el tiempo de entrenamiento. Así pues, dos capas ocultas de 100 y 50 neuronas respectivamente parece una estructura óptima. Sin embargo, no parece ser única, ya que se pueden alcanzar los mismos resultados con diferentes estructuras.

6.3.2. Número de neuronas por capa

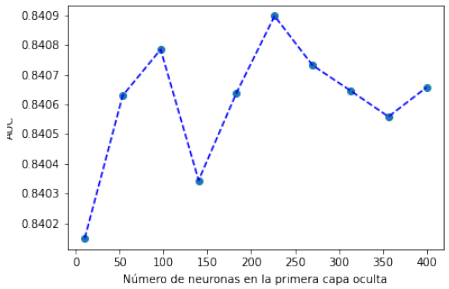
Por otro lado, se estudiará la variación del comportamiento de la red en función del número de neuronas en cada capa del modelo.

En primer lugar, se comenzará modificando el número de neuronas en la primera capa oculta. Hasta el momento, se habían utilizado 100 neuronas. En este estudio se consideraron 10 cifras equiespaciadas comprendidas entre 10 y 400. La red utilizada consta, además, de una segunda capa oculta de 50 neuronas y una capa de salida con únicamente una neurona. En la Figura 6.4 se muestra la alteración del error de validación cruzada y el AUC de la red al cambiar el número de neuronas de la primera capa oculta. En las gráficas, se observa muy poca variación tanto en el error de validación como en el AUC.

Por tanto, a partir de las representaciones gráficas de 6.4, se puede concluir que aumentar el número de neuronas en la primera capa oculta no mejora el rendimiento de la red. Además, cuántas más neuronas se utilicen, mayor será el tiempo de entrenamiento de la red, llegando incluso a duplicarse.



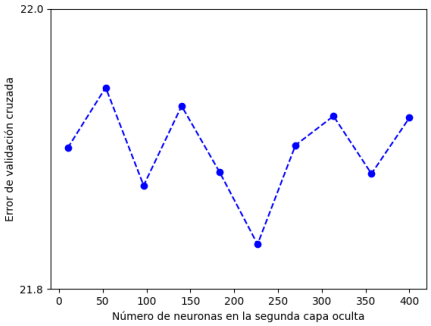
(a) Variación del error de validación curzada de la red neuronal frente al número de neuronas en la primera capa oculta de la red.



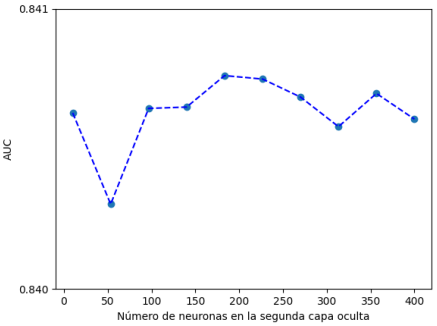
(b) Variación del AUC de la red neuronal frente al número de neuronas en la primera capa oculta de la red

Figura 6.4: Variación del rendimiento de la red neuronal en función del número de neuronas en la primera capa oculta.

Del mismo modo, se estudió la variación de los resultados de la red en función del número de neuronas de la segunda capa oculta, manteniendo la capa inicial con 100 neuronas.



(a) Variación del error de validación curzada de la red neuronal frente al número de neuronas en la segunda capa oculta de la red.



(b) Variación del AUC de la red neuronal frente al número de neuronas en la primera capa oculta de la red

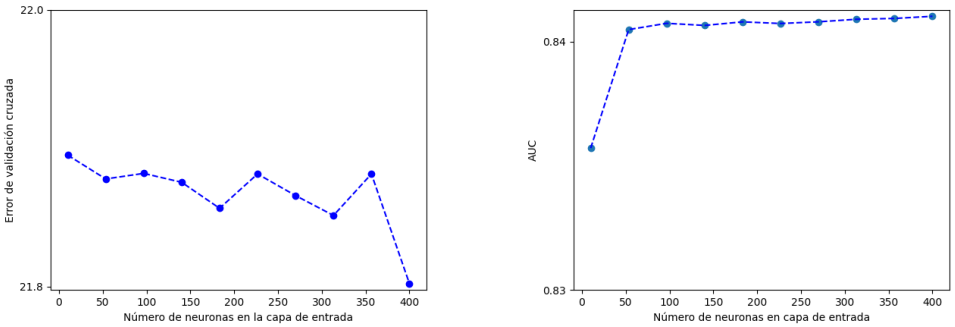
Figura 6.5: Variación del rendimiento de la red neuronal en función del número de neuronas en la segunda capa oculta.

Al igual que en el primer caso, ni el aumento ni la disminución del número de neuronas en la segunda capa oculta de la red provocan un cambio en el rendimiento del modelo. Dado que, a mayor número de neuronas, mayor coste computacional, parece óptimo escoger un número entre 10 y 100.

En resumen, aumentar el número de neuronas en las capas ocultas no mejora el rendimiento de la red, pero aumenta el tiempo de entrenamiento. Como consecuencia, la convergencia de

modelo parece independiente de la arquitectura del mismo. De hecho, una arquitectura sencilla es capaz de encontrar el mínimo local y converger.

Finalmente, se analizó la influencia del número de neuronas de la capa de entrada de la red en los resultados de la misma. De forma análoga, se consideraron 10 cifras equiespaciadas comprendidas entre 10 y 400. Los resultados obtenidos se muestran en la Figura 6.6. Al igual que antes, el rendimiento de la red parece independiente del número de neuronas en la capa de entrada. Sin embargo, cabe destacar que cuando se utilizan únicamente 10 neuronas de entrada, el AUC del modelo baja a 0.835. Como consecuencia, parece que el modelo requiere un número mínimo de neuronas en la capa de entrada para alcanzar un rendimiento óptimo.



(a) Variación del error de validación curzada de la red neuronal frente al número de neuronas en la capa de entrada de la red.

(b) Variación del AUC de la red neuronal frente al número de neuronas en la capa de entrada de la red

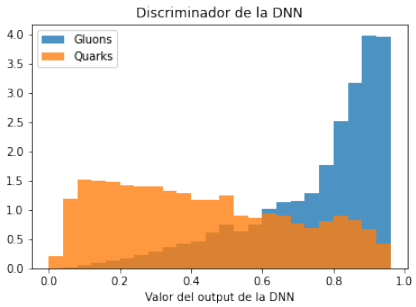
Figura 6.6: Variación del rendimiento de la red neuronal en función del número de neuronas en la capa de entrada.

6.4. Variable de salida de la red neuronal

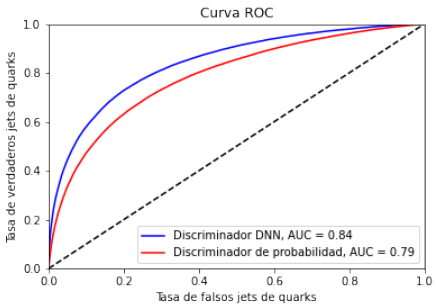
La variable de salida de la red es una variable que toma valores entre 0 y 1, siendo 0 cuando el origen del jet es un gluon y 1 cuando es un quark ligero. Como se ha comentado anteriormente, la red neuronal comete menor error al predecir los jets originados por gluones frente a aquellos originados por un quark ligero. De hecho, la red neuronal etiqueta correctamente el 87 % de los jets originados por gluones frente al 63 % de los jets de quarks ligeros. Esto puede deberse a que en la muestra, existe un mayor número de jets originados por gluones, en concreto hay aproximadamente un 60 % más de jets originados por gluones que por quarks ligeros.

Aún así, cabe preguntarse si se obtendrían los mismos resultados utilizando como variable de salida aquella que, por el contrario, tome el valor 1 cuando el origen del jet sea un gluon y

0 en caso de los quarks ligeros. Así, se ha contruido una red neuronal equivalente cuya única diferencia es la variable de salida.



(a) Distribución del discriminador construido con la red neuronal profunda.



(b) Representación gráfica de la curva ROC para la red neuronal profunda y para el discriminador de probabilidad. Se obtienen respectivamente unos valores de AUC de 0.84 y 0.79.

Figura 6.7: Resultados obtenidos utilizando como variable de salida de la red neuronal aquella que asigna el valor 0 en el caso de jets originados por quarks ligeros y 1 en el caso de los gluones.

En la Figura 6.7, se presentan la distribución del discriminador y la curva ROC obtenidas. Destaca que, en este caso, la distribución del discriminador para los jets originados por gluones alcanza el máximo en 1 mientras que la de los quarks lo hace cerca de 0, de acuerdo con lo esperado. Además, la matriz de confusión obtenida fue la siguiente:

$$\begin{pmatrix} 0.24 & 0.14 \\ 0.08 & 0.54 \end{pmatrix} \tag{6.4}$$

En este caso, la primera fila de la matriz se corresponde con los jets generados por quarks.

Asimismo, la precisión para los jets originados por quarks ligeros fue de 0.75 frente a 0.80 de los gluones, luego la red sigue etiquetando mejor los jets originados por gluones. Por otro lado, la exhaustividad alcanzó los valores de 0.64 y 0.87 respectivamente, mejorando en un 1 % la fracción de jets de quarks ligeros clasificados correctamente. Finalmente, el error de validación cruzada del modelo fue de 22 %, al igual que utilizando la variable de salida contraria.

En resumen, al invertir la variable de salida de la red neuronal, optando por aquella que asigne el valor 1 a los jets originados por gluones y 0 a los de quarks ligeros, los resultados se mantienen prácticamente inalterados. Esto es, por tanto, una prueba más de la robustez del modelo.

Conclusiones

Las redes neuronales profundas han resultado ser una herramienta muy útil en el análisis de los datos, permitiendo diferenciar jets originados por gluones o quarks ligeros. Las colisiones de partículas son procesos muy complejos y cuyos cálculos pueden ser muy enrevesados. El aprendizaje automático y, en particular, las redes neuronales profundas, ofrecen una alternativa más sencilla para la comprensión de estos fenómenos. Además, destaca la robustez del modelo, ya que se ha comprobado, que converge independientemente de su arquitectura.

Al mismo tiempo, la red neuronal ha demostrado un rendimiento notablemente superior en relación a los resultados previos, alcanzados por el discriminador de probabilidad. Con el uso de la red neuronal, se logró clasificar la muestra *test* con una precisión de 0.78. Además, el área bajo la curva ROC alcanzado fue de 0.84, superando la cifra de 0.79, obtenida con el discriminador de probabilidad.

Ahora bien, se ha constatado que el rendimiento de la red neuronal se ve afectado por los observables suministrados. De hecho, propocionar los observables óptimos a la red neuronal, ha resultado ser crucial en la mejora de la clasificación. Como consecuencia, para poder alcanzar los mejores resultados, es necesaria una comprensión profunda del problema. Con esta finalidad, a lo largo del trabajo se ha explorado el funcionamiento del experimento CMS dentro del LHC y la física de los jets.

Igualmente, la red neuronal ha resultado ser más eficiente en el etiquetado de gluones. La razón puede encontrarse en que se ha trabajado con una muestra descompensada, con un 60 % más de jets originados por gluones, debido a que su radiación es más probable en QCD.

Por otro lado, cabe mencionar que, debido a limitaciones computacionales, sólo se utilizó el 19 % de la información disponible. El uso de un dispositivo más potente permitiría mejorar los resultados alcanzados en este trabajo.

Bibliografía

[1] **CERN** Collaboration, "The Standard Model," tech. rep., CERN, Visitado en septiembre 2023. https://home.cern/science/physics/standard-model.

[2] J. Stone, "Did the Universe Just Break up with Us?,". http://www.thepromptmag.com/did-the-universe-just-break-up-with-us/. Tabla de partículas tomada el 01/03/2023.

[3] R. Nave, "Fundamental forces," *Tabla de fuerzas fundamentales* . http://hyperphysics.phy-astr.gsu.edu/hbase/Forces/funfor.html. Visitado el 29/11/2023.

[4] M. Strassler, "Conversations about science with theoretical physicist matt strassler," *Figure 2* (2013) . https://profmattstrassler.com/articles-and-posts/particle-physics-basics/the-structure-of-matter/protons-and-neutrons/. Visitado en septiembre 2023.

[5] A. Martin, W. Stirling, R. Thorne y G. Watt., "Parton distributions for the LHC," *Eur.Phys.J.C63:189-285,2009* (2009) 5. https://doi.org/10.48550/arXiv.0901.0002.

[6] V. Buxbaum, "Cross section and luminosity," *Physics Cheat Sheet* (2022) 1. https://cds.cern.ch/record/2800984.

[7] **CMS** Collaboration, "Cms luminosity public plots," tech. rep., CERN. https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults.

[8] **CMS** Collaboration, "How CMS weeds out particles that pile up," tech. rep., CERN, Visitado en enero 2024. https://cms.cern/news/how-cms-weeds-out-particles-pile.

[9] A. Babaev, "Public CMS luminosity information," *Interactions per crossing (pileup)* (2023) . https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults.

[10] **CERN** Collaboration, "The Large Hadron Collider," tech. rep., CERN, Visitado en septiembre 2023. https://home.cern/science/accelerators/large-hadron-collider.

[11] E. Lopienska, "The CERN accelerator complex," tech. rep., CERN, 2022. https://cds.cern.ch/images/CERN-GRAPHICS-2022-001-1.

[12] **CMS** Collaboration, "Experiments," tech. rep., CERN, Visitado en diciembre 2023. https://home.cern/science/experiments.

[13] **CMS** Collaboration, "The CMS Experiment at CERN: Detector," tech. rep., CERN, Visitado en septiembre 2023. https://cms.cern/detector.

[14] Lippmann, C, "Particle identification," (2011) 2. https://doi.org/10.48550/arXiv.1101.3276.

[15] **CMS** Collaboration, "Detecting muons," tech. rep., CERN, Visitado el 03/12/2023. https://cms.cern/index.php/detector/detecting-muons.

[16] A.M. Sirunyan et al, "Particle-flow reconstruction and global event description with the CMS detector," *JINST 12 (2017) P10003* **2** (2017) . https://doi.org/10.48550/arXiv.1706.04965.

[17] I. Neutelings, "CMS coordinate system,". https://tikz.net/axis3d_cms/. Visitado el 03/12/2023.

[18] **CMS** Collaboration, "Triggering and data acquisition," tech. rep., CERN, Visitado el 03/12/2023. https://cms.cern/detector/triggering-and-data-acquisition.

[19] C. Bierlich, et al, "A comprehensive guide to the physics and usage of PYTHIA 8.3" (2022) 7–16. https://doi.org/10.48550/arXiv.2203.11601.

[20] T. Potter, "Qcd: Particle and nuclear physics," *QCD* (2023) 1–14. https://www.hep.phy.cam.ac.uk/~chpotter/particleandnuclearphysics/mainpage.html. Visitado en septiembre 2023.

[21] **ATLAS** Collaboration, "Flavour Tagging with Graph Neural Networks @ ATLAS," tech. rep., CERN, 2023. https://indico.cern.ch/event/1232499/attachments/2602341/4494127/2023-03-01_GN1_Seminar.pdf.

[22] **CMS** Collaboration, "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV," tech. rep., CERN, 2017. https://cms-results.web.cern.ch/cms-results/public-results/publications/BTV-16-002/.

[23] M. Tasevsky, "Differences between quark and gluon jets as seen at lep," (2001) 1–8. https://doi.org/10.48550/arXiv.hep-ex/0110084.

[24] J. Gary, "Determination of the QCD color factor ratio CA/CF from the scale dependence of multiplicity in three jet events," *Phys.Rev. D61 (2000) 114007* (1999) 1–6. https://doi.org/10.48550/arXiv.hep-ex/9911011.

[25] **CMS** Collaboration, "Performance of quark/gluon discrimination in 8 TeV pp data," tech. rep., CERN, Geneva, 2013. https://cds.cern.ch/record/1599732.

[26] E. Alpaydin, *Introduction to Machine Learning*. No. 3. MIT Press, 2014.

[27] **Particle Data Group** Collaboration, K. Crammer, U. Seljak y K.Terao, "Machine learning," *Prog.Theor.Exp.Phys.2022, 083C01 (2022)* 3–11, 21–47.

[28] Google Developers, "Google machine learning crash course: Roc and auc,". https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=es-419. Visitada el 07/12/2023.

[29] Google Developers, "Google machine learning crash course: Precision and recall,". https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=es-419. Visitada el 07/12/2023.

[30] H2O.ai, "Overfitting," (2023) . https://h2o.ai/wiki/overfitting/. Visitada el 06/12/2023.

[31] J. Baek and Y. Choi, "Deep Neural Network for Predicting Ore Production by Truck-Haulage Systems in Open-Pit Mines," (2020) 5. https://doi.org/10.3390/app10051657. Figure 4(a).

[32] H. Yadav, "Dropout in Neural Networks," (2022) . https://towardsdatascience.com/dropout-in-neural-networks-47a162d621d9. Visitada el 08/12/2023.

[33] K. Kallonen, "Sample with jet properties for jet-flavor and other jet-related ML studies," tech. rep., CERN Open Data Portal, 2019. https://doi.org/10.7483/OPENDATA.CMS.RY2V.T797.

[34] S. Agostinelli et al., *Geant4 - A Simulation Toolkit*, pp. 250–303. Nucl. Instrum. Meth., 2003. https://doi.org/10.1016/S0168-9002(03)01368-8.

[35] F. Beaudette, "The CMS Particle Flow Algorithm," *Proceedings of the CHEF2013 Conference - Eds. J.C. Brient, R. Salerno, and Y. Sirois - p295 (2013), ISBN 978-2-7302-1624-1* (2018) 2–4. https://doi.org/10.48550/arXiv.1401.8155.

[36] M. C. y G. Salam, "The anti-kt jet clustering algorithm," *JHEP 0804:063,2008* **2** 1–5. https://doi.org/10.48550/arXiv.0802.1189.

[37] T. Carpentries, "CMS Jets and MET: Pile up," (2020) . https://cms-opendata-workshop.github.io/workshop-lesson-jetmet/01-jetmet/index.html. Visitado en diciembre 2023.

[38] **CMS Open Data Guide** Collaboration, "Jet Energy Corrections (JEC)," tech. rep., CERN, Visitado el 08/01/2024. https://cms-opendata-guide.web.cern.ch/analysis/systematics/objectsuncertain/jetmetuncertain/.

[39] **CERN Open Data Portal** Collaboration, "CMS Monte Carlo production overview," tech. rep., CERN, Visitado en enero 2024. https://opendata.cern.ch/docs/cms-mc-production-overview.

[40] J. Gallicchio y M. D. Schwartz, "Quark and Gluon Tagging at the LHC," *Phys.Rev.Lett. 107 (2011) 172001* 1–5. https://doi.org/10.48550/arXiv.1106.3076.

[41] A. Sirunyan, "Particle-flow reconstruction and global event description with the cms detector," *JINST 12 (2017) P10003* . https://doi.org/10.48550/arXiv.1706.04965.

[42] N. Walet, "The feynman diagrams of QCD," (2011) . https://oer.physics.manchester.ac.uk/NP/Notes/Notes/Notesse54.xht. Visitado en diciembre 2023.

[43] J. Álvarez Rodríguez, "Caracterización de los algoritmos de etiquetado de quarks b en CMS," *Trabajo fin de grado* (2019) . Universidad de Oviedo.

Apéndice A

Código

El código representa una parte importante de este trabajo. Para desarrollar la red neuronal profunda, se empleó el lenguaje de programación *Python*. En particular, se trabajó con las bibliotecas *Pandas*, *TensorFlow*, *Keras* y *sklearn.metrics*. *Pandas* facilita la manipulación eficiente de grandes conjuntos de datos. Por otro lado, *TensorFlow* posibilita la creación sencilla de modelos de aprendizaje automático, mientras que *Keras* se especializa en redes neuronales profundas. Finalmente, *sklearn.metrics* implementa funciones que simplifican cálculos relacionados con problemas de clasificación.

Además, el trabajo se ha llevado a cabo en la plataforma *Kaggle*[1], una herramienta de trabajo diseñada por Gooogle para la implementación de modelos de inteligencia artifical y análisis de datos. Destaca por proporcionar un amplio espacio de almacenamiento en la nube, grandes cantidades de CPU y una memoria RAM de 30GB. Además, permite el uso de hasta 30 horas semanales de GPU (T4 x2 y P100), que permiten acelerar el procesamiento de la red neuronal profunda.

En nuestro caso, el tiempo de ejecución era aproximadamente 20 minutos, del orden de un minuto para cada etapa o época del entrenamiento.

El código que he realizado para este trabajo está en el siguiente repositorio:

`https://github.com/Rociocoro7/Codigo-TFG.git`

[1] `https://www.coursera.org/articles/kaggle`