Estudio con aprendizaje automático de la certificación de datos físicos del detector CMS de LHC (CERN) utilizando Análisis de Componentes Principales (PCA)

Manuel Iglesias Alonso

Tutores:

Javier Fernández Menéndez

Andrea Trapote Fernández

Grado en Física

Universidad de Oviedo

Junio 2023

Prólogo

ἐτεῇ δὲ ἄτομα καὶ κενόν
En realidad solo átomos y vacío
-Demócrito

Guiados por la curiosidad inherente al ser humano y en un intento por explicar la naturaleza del mundo que nos rodea, surgió en el siglo V a.C. una corriente filosófica de la Antigua Grecia conocida en la actualidad como atomismo y cuyo máximo exponente es Demócrito. Proponían que la materia no podía ser dividida de forma indefinida, sino que llegado un punto se encontraría un componente fundamental e indivisible al que llamaron ἄτομον, átomo.

La idea atómica fue desplazada a un segundo plano durante más de dos milenios debido a la popularidad de la teoría propuesta por Aristóteles para la materia, basada en los cuatro elementos: tierra, fuego, aire y agua. Estos serían continuos y no formados por unidades indivisibles. El escepticismo en torno a estas ideas cobró fuerza con la Revolución Científica de los siglos XVI y XVII, donde la noción de elemento químico sustituyó a los elementos aristotélicos.

La teoría atómica reaparece de la mano de John Dalton en los primeros años del siglo XIX, manteniendo la idea clásica de la indivisibilidad propuesta por los atomistas. Una vez más, estas ideas no recibieron una aceptación inmediata y fue necesario el transcurso de varias décadas para su consolidación gracias a la formalización teórica y su capacidad para explicar las propiedades periódicas observadas para los elementos químicos.

La noción del átomo como unidad fundamental se vería dinamitada en 1897 con el descubrimiento del electrón por JJ Thomson, lo que abría la puerta a considerar la existencia de partículas subatómicas, reforzado por los descubrimientos del protón y neutrón en las siguientes décadas. Estos años también se corresponden con el nacimiento de la teoría cuántica.

La primera mitad del siglo XX vive el descubrimiento de nuevas partículas gracias al uso de detectores relativamente sencillos que permiten estudiar fenómenos naturales como los rayos cósmicos. El interés por ser capaz de obtener estas nuevas partículas de forma controlada lleva a la creación de los primeros dispositivos que permiten acelerar partículas cerca de la velocidad de la luz para generar colisiones muy energéticas en las que poder producirlas.

El esfuerzo colectivo por mejorar las técnicas de aceleración, detección y estudio de estas colisiones ha permitido el descubrimiento de cientos de partículas en las últimas décadas y el desarrollo de un marco teórico que resume nuestro conocimiento actual sobre ello, conocido como el modelo estándar. En la actualidad el máximo exponente es el acelerador LHC, perteneciente a la organización internacional conocida como CERN. Este trabajo se centra en uno de los detectores del mismo, conocido como CMS.

El funcionamiento de CMS involucra la colaboración de más de 5000 personas y permite recoger un enorme volumen de datos que debe ser procesado y analizado mediante una de las obras tecnológicas más avanzadas de la historia. En un sistema de obtención de datos tan complejo es necesaria una estricta supervisión para garantizar la calidad de los mismos para la búsqueda de nueva física. Así, la certificación consiste en la clasificación de los datos recogidos como buenos o malos y en la actualidad se basa en la revisión manual de los mismos por expertos humanos. Este trabajo se ha vuelto cada vez más complejo debido a la creciente cantidad de información con la que se trabaja, llevando a la necesidad de proponer técnicas de automatización.

Este trabajo propone un modelo de aprendizaje automático para la certificación de datos en CMS basado en el análisis de componentes principales. Este será entrenado y evaluado con datos recogidos en los últimos años en este detector pero presenta un importante potencial para su aplicación en el futuro.

El trabajo comenzará con tres capítulos que presentarán los fundamentos de la física de altas energías y sus experimentos, el acelerador LHC, el detector CMS y el aprendizaje automático. Esto resultará esencial para comprender la relevancia del problema planteado y las bases sobre las que se asienta. El cuarto capítulo se centra en la introducción de la idea clásica del análisis de componentes principales, así como la propuesta de un modelo basado en esta técnica y que permite su uso en la certificación de datos físicos. El trabajo finaliza con un capítulo destinado a evaluar el comportamiento del modelo creado, así como un estudio detallado de los valores óptimos de los parámetros de los que depende.

Índice general

Índice de figuras

Capítulo 1

Introducción a la física de altas energías

1.1. El modelo estándar

La física de partículas se encarga del estudio de los constituyentes más pequeños de la materia y de las interacciones entre ellos. Es en este contexto en el que surge la idea de partícula elemental, entendida como aquella que no puede ser dividida, es decir, que no está formada por otros componentes más pequeños [1, 2].

El modelo estándar es un conjunto de teorías que explica nuestro conocimiento actual de las partículas elementales y las interacciones entre ellas. Se fundamenta en la idea de que existen dos tipos de partículas: los fermiones, con spin semientero y que son los constituyentes fundamentales de la materia, y los bosones, con spin entero y responsables de que se produzcan los distintos tipos de interacciones. En la actualidad se conocen un total de doce fermiones elementales y sus correspondientes antipartículas (tienen todas las propiedades idénticas a excepción de la carga eléctrica, que es de signo contrario), así como cinco bosones elementales [3].

Existen cuatro interacciones fundamentales: gravitatoria, electromagnética, fuerte y débil. Sin embargo, la gravitatoria no es relevante a nivel subatómico por ser mucho menos intensa que las demás. Estas se explican por medio del intercambio de partículas con spin entero, llamadas bosones: el fotón para la interacción electromagnética, el gluon para la fuerte y los bosones $W^{\pm}$ y Z para la débil (todos estos tienen spin 1). A cada interacción se le asocia una carga (eléctrica, fuerte o débil) y únicamente aquellas partículas que la tengan experimentarán la correspondiente interacción. En la actualidad, las interacciones electromagnética y débil se pueden describir como manifestaciones de una interacción unificada conocida como fuerza electrodébil.

Los fermiones elementales se van a dividir en dos grupos, llamados quarks y leptones, todos estos de spin 1/2 y que se diferencian en el hecho de que los primeros experimentan la interacción fuerte y los segundos no. Todos ellos poseen carga débil y, a excepción de las partículas llamadas neutrinos, también tienen carga eléctrica. Un hecho interesante es que el gluon posee carga fuerte y a la vez es el portador de la correspondiente interacción, algo que no ocurre en el caso del fotón, que no tiene carga eléctrica.

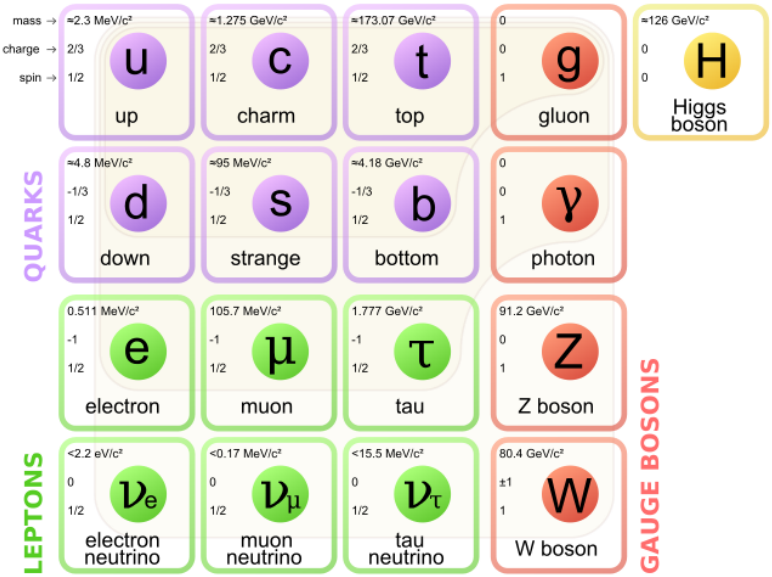


Figura 1.1: Resumen de las partículas del Modelo Estándar. Fuente: [4]

Además de los bosones previamente mencionados existe uno adicional llamado bosón de Higgs. Esta es una partícula sin carga eléctrica y con spin 0 que interacciona con las demás permitiendo explicar un mecanismo para que estas adquieran masa. Fue propuesta de forma teórica en la década de 1960 y ha sido la última partícula confirmada experimentalmente a partir del análisis de los datos recogidos en 2011 y 2012 por los detectores ATLAS y CMS del LHC (CERN).

Una partícula que va a tener especial relevancia en este trabajo es el muon, representado por la letra griega μ. Este es idéntico al electrón a excepción de su masa, que es más de 200 veces superior, y su tiempo de desintegración, ya que mientras el electrón es estable, el muon tiene una vida media de 2.2 μs. Este valor, si bien es pequeño, es muy superior al del resto de partículas inestables, lo que va a resultar clave en los experimentos para su detección, ya que va a permitir identificar los muones a una distancia grande del punto en el que sean producidos.

1.2. Experimentos

Al igual que ocurre en muchas otras áreas de la física, los experimentos juegan un papel fundamental en la física de altas energías para comprobar la validez de los modelos existentes para explicar el universo, poniendo a prueba el modelo estándar. Además tienen una gran importancia de cara a determinar parámetros de dichos modelos.

Estos experimentos requieren en la mayor parte de los casos del uso de aceleradores de partículas, ya que permiten disponer de haces de partículas con energías muy elevadas que se hacen colisionar contra un blanco fijo o entre ellos, lo que permite obtener estados excitados o nuevas partículas. La equivalencia entre masa y energía dada por la conocida fórmula $E = mc^2$ es la responsable de que sean necesarias energías muy elevadas en dichas colisiones para poder generar nuevas partículas mediante dicha conversión.

Se distinguen fundamentalmente dos tipos de experimentos para producir colisiones:

- **Experimentos de blanco fijo:** Se acelera un haz de partículas y se hace colisionar este con un blanco estacionario. Los aceleradores lineales son dispositivos que permiten realizar este tipo de experimentos, consistiendo en una serie de tubos acoplados en línea recta (tubos de deriva) que van alternando su polaridad con la frecuencia adecuada para que las partículas cargadas que circulan en su interior sean siempre aceleradas debido a las sucesivas diferencias de potencial. Este sistema no permite producir un flujo continuo, ya que las partículas estarán agrupadas formando paquetes.

  Otra posibilidad para este tipo de experimentos son los sincrotrones con un único haz. En ese caso las partículas se mantienen en una trayectoria circular gracias a un campo magnético que aumenta a medida que crece la velocidad de las mismas y son aceleradas repetidamente por una diferencia de potencial localizada en algún punto de su recorrido.

  Uno de los parámetros más importantes de un acelerador es la energía en centro de masas, definida como la energía total en el sistema de referencia del centro de masas, aquel en el que el momento total de las partículas que colisionan es 0. Para un experimento de blanco fijo viene dada por:

  $$\sqrt{s} = \sqrt{2E_a m_b c^2}$$

  Aquí E_a es la energía del haz y m_b es la masa del blanco fijo con el que se hace colisionar.

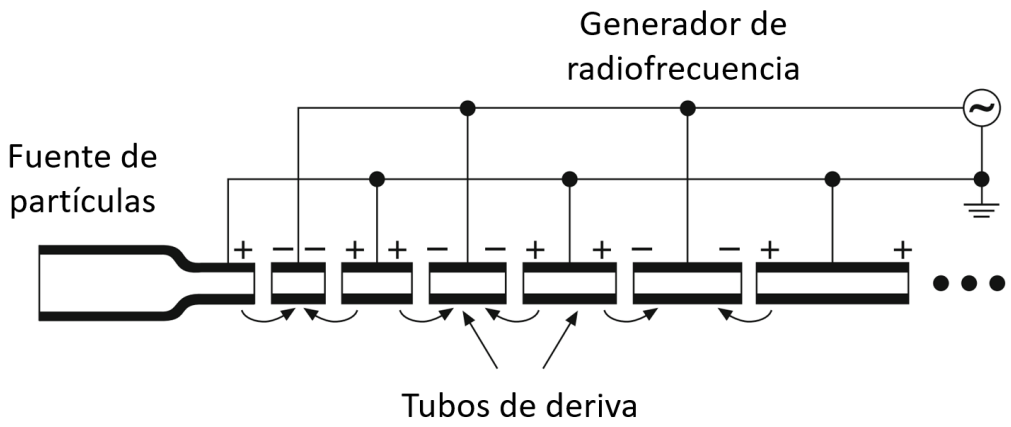


Figura 1.2: Esquema de un acelerador lineal. Fuente: Modificado de [2]

- **Colisionadores:** Permiten acelerar dos haces de partículas en sentidos opuestos y posteriormente los hacen colisionar entre sí. Al igual que ocurría con los experimentos de blanco fijo, pueden ser aceleradores lineales o circulares. En este caso, la energía en centro de masas es:

$$\sqrt{s} = 2E_a$$

  Aquí E_a vuelve a ser la energía de cada uno de los haces de partículas. Este valor de $\sqrt{s}$ es muy superior al obtenido mediante experimentos de blanco fijo, siendo esta la principal ventaja que justifica su utilización, mientras que su principal desventaja es que se reduce la tasa de colisiones. Los datos que se emplean en este trabajo fueron recogidos por el detector CMS en el colisionador LHC del CERN, como se describe en detalle en el capítulo 2.

1.3. Detectores

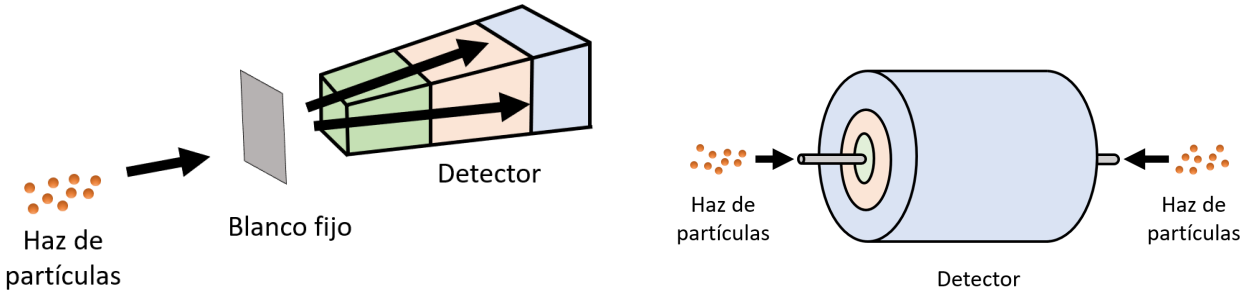
La detección de las partículas resultantes de las colisiones se realiza fundamentalmente a partir del estudio de su interacción con la materia. La pérdida de energía de las partículas cargadas se debe principalmente a las interacciones con las nubes electrónicas de los átomos, provocando que estos se ionicen o pasen a estados excitados. Esta pérdida, descrita por la fórmula de Bethe-Bloch, depende de la carga y velocidad de las partículas que se están detectando y de las propiedades del medio que atraviesan, siendo estas últimas conocidas en el caso de un detector. En el caso de los electrones y positrones, debido a su pequeña masa, una causa muy importante de pérdida de energía en su interacción con la materia es el *bremsstrahlung*, que es la radiación emitida por estas partículas cuando son frenadas.

En el caso de los fotones, la interacción con la materia no se produce de forma progresiva como en otras partículas, sino que ocurre en un punto y tiene lugar mediante tres procesos: efecto fotoeléctrico, efecto Compton y producción de pares. El efecto fotoeléctrico se produce cuando el fotón es absorbido por un electrón de un átomo dando lugar a su emisión, el efecto Compton es la dispersión de un fotón por una colisión con un electrón y la producción de pares es la creación de un par electrón-positrón. La dominancia de uno u otro método de interacción depende del rango de energía del fotón, siendo el efecto fotoeléctrico el más relevante a bajas energías, el efecto Compton a energías intermedias y la producción de pares para fotones muy energéticos.

La estructura y componentes de un detector va a depender de su finalidad concreta y de los experimentos en los que participe, si bien es habitual que se emplee una estructura en capas en la que cada una tenga el objetivo de estudiar unas ciertas propiedades o un tipo de partículas.

Una capa se destina generalmente a medir la posición de las partículas, por ejemplo mediante la ionización de un gas o la generación de pares electrón-hueco producidos por el paso de una partícula cargada eléctricamente. Para medir el momento y la carga eléctrica se usa un campo magnético, de manera que la trayectoria de las partículas cargadas se curva dando lugar a un recorrido circular cuyo radio depende de p y su sentido de giro del signo de la carga. La medición de la energía de la partícula se basa habitualmente en que sea absorbida totalmente por el medio. Para ello se emplea instrumentación como detectores semiconductores y calorímetros.

La geometría de un detector depende del tipo de experimento, adaptándola para minimizar la pérdida de información. En un experimento de blanco fijo el detector se suele situar detrás de este y cubriendo los posibles ángulos de dispersión, mientras que en un colisionador se utiliza una geometría cilíndrica al no existir una dirección preferencial de las partículas producidas.



(a) Experimento de blanco fijo. (b) Colisionador.

Figura 1.3: Geometría de los detectores. Fuente: Elaboración propia.

El momento lineal $\mathbf{p}$ de una partícula se puede expresar a través de sus componenentes en coordenadas cartesianas:

$$\mathbf{p} = (p_x, p_y, p_z)$$

En este trabajo nos vamos a centrar en los detectores cilíndricos de los colisionadores, en concreto en CMS, lo que lleva a considerar un nuevo sistema de coordenadas:

$$\mathbf{p} = (p_t, \eta, \varphi)$$

A continuación vamos a describir brevemente cada una de estas magnitudes:

- $\mathbf{p_t}$: Es el momento transverso de la partícula. Como su nombre indica, se corresponde con la componente transversal del momento lineal, es decir, situada en el plano perpendicular a la dirección del haz.

- $\boldsymbol{\eta}$: Recibe el nombre de pseudorapidez. Se define como sigue:

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

  Aquí θ se corresponde con el ángulo polar en coordenadas esféricas. La pseudorapidez toma valores entre $-\infty$ e $+\infty$ y es una buena aproximación a la rapidez, una magnitud más difícil de calcular y que tiene invariancia ante ciertas transformaciones Lorentz [5].

- $\boldsymbol{\varphi}$: Es el ángulo azimutal en coordenadas esféricas y cilíndricas.

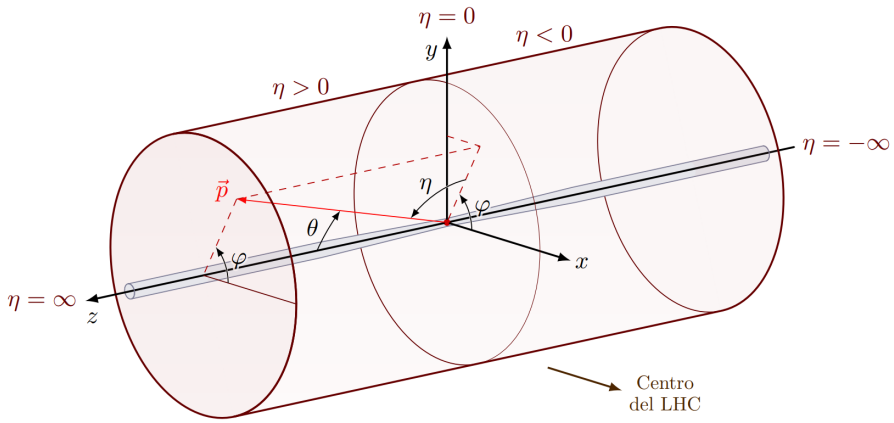


Figura 1.4: Sistema de coordenadas en un detector cilíndrico. Fuente: Modificado de [6]

Capítulo 2

El Gran Colisionador de Hadrones

2.1. LHC

El CERN (Organización Europea para la Investigación Nuclear, siglas en francés) es un laboratorio de física de partículas localizado principalmente en Ginebra y que basa su funcionamiento en la colaboración internacional para llevar a cabo investigación en física fundamental.

El CERN presenta un sistema de aceleradores encadenados, de manera que cada uno de ellos inyecta el haz de partículas en el siguiente y lo hacen de forma que su energía vaya aumentando de forma progresiva, siendo el Gran Colisionador de Hadrones (LHC) el último eslabón. La mayor parte de estos aceleradores previos tienen sus propios experimentos asociados para los que se necesita una energía inferior que la alcanzada al final de la cadena [7, 8].

La mayor parte de las colisiones en el LHC se realizan entre haces de protones, si bien existen periodos más cortos de funcionamiento en los que también se estudian colisiones entre núcleos de plomo. En el periodo actual de funcionamiento, denominado *Run* 3, se puede alcanzar una energía de 6.8 TeV para cada haz de protones, lo que supone una energía en centro de masas de 13.6 TeV, el valor más alto alcanzado hasta la fecha.

Además de estos experimentos, también existen otros que no basan su investigación en los aceleradores de partículas, como puede ser AMS, un detector controlado desde el CERN pero que se encuentra en un módulo acoplado a la Estación Espacial Internacional para la búsqueda de materia oscura y antimateria.

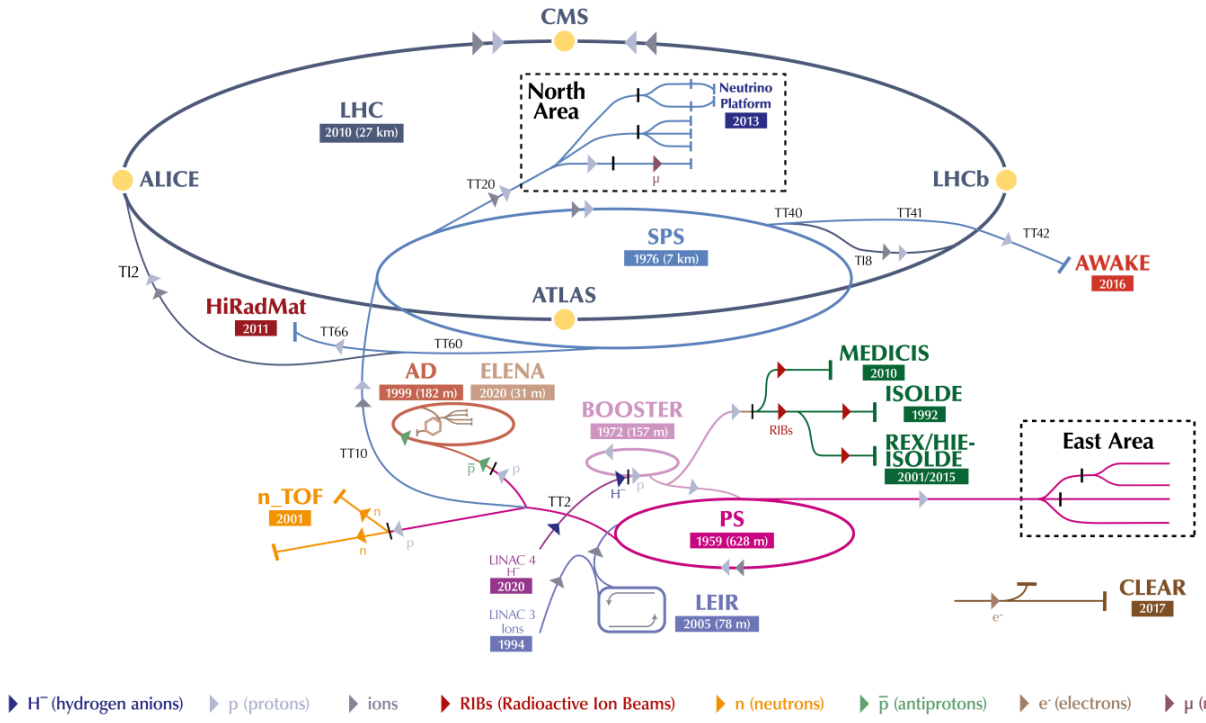


Figura 2.1: Conjunto de aceleradores encadenados del CERN. Fuente: [7]

Como se ha comentado previamente, la aceleración de los protones con el fin de conseguir la máxima energía posible en las colisiones se realiza por etapas mediante una cadena de aceleradores que se representa en la figura 2.1. El procedimiento es el siguiente:

1. Los protones utilizados en las colisiones se obtienen a partir de iones de hidrógeno H^-, que son acelerados hasta 160 MeV mediante el acelerador lineal LINAC 4. La pérdida de los dos electrones se produce durante la inyección de estos átomos en la siguiente etapa, el Booster, quedando únicamente protones para las sucesivas fases de aceleración.

2. El Booster es un sincrotrón que permite acelerar los protones hasta 2 GeV.

3. El Sincrotrón de Protones (PS) aumenta la energía de los protones a 26 GeV. Este es el primer acelerador común a protones e iones de plomo.

4. Una vez alcanzada dicha energía se introducen en el Súper Sincrotrón de Protones (SPS), donde llegan a 450 GeV. Este sincrotrón cuenta con varios experimentos a menores energías que el LHC.

5. El último paso es la inyección en el LHC, donde se dividen en dos haces que giran en sentidos opuestos.

El Gran Colisionador de Hadrones es el acelerador de partículas más grande y con mayor energía de colisión del mundo. Presenta una circunferencia de 27 kilómetros de longitud y discurre a una profundidad de entre 50 y 200 metros bajo los territorios de Suiza y Francia. Presenta dos tuberías circulares concéntricas por las que circulan los dos haces de partículas en sentidos opuestos y que se intersecan en cuatro puntos del recorrido para que se produzcan las colisiones, lugares en los que se encuentran los cuatro detectores principales de LHC.

La aceleración de los protones tiene lugar en regiones localizadas del anillo conocidas como cavidades de radiofrecuencia, en las cuales son sometidos a una diferencia de potencial que los acelera. Este potencial oscila con una frecuencia que busca garantizar que los protones se mantengan agrupados en paquetes conocidos como *bunches*. Para ello, la frecuencia es tal que una vez que han alcanzado la energía adecuada para la colisión, aquellos protones que estén correctamente situados no experimentarán más aceleración, mientras aquellos que se encuentren desfasados serán acelerados o frenados para que se mantengan unidos al grupo.

Los electroimanes superconductores constituyen también unos elementos fundamentales para el correcto funcionamiento del colisionador, cumpliendo una amplia variedad de funciones:

- **Curvar la trayectoria:** Más de mil imanes dipolares dispuestos a lo largo del LHC generan campos de hasta 8T para curvar los protones y que puedan describir la trayectoria circular necesaria. Es una de las limitaciones técnicas del LHC, pues conseguir energías de colisión superiores requiere aumentar el radio del acelerador o el valor de estos campos.
- **Colimación:** Electroimanes cuadrupolares cumplen la función de mantener un haz de protones muy colimado para aumentar la probabilidad de colisión, ajustando su posición en los ejes horizontal y vertical.
- **Inyección en detectores:** Dos tripletes de imanes cuadrupolares se sitúan a la entrada de cada detector con la finalidad de colimar aún más el haz cuando se van a producir las colisiones en su interior.
- **Medición:** Los detectores incluyen potentes imanes con el fin de curvar las trayectorias de las partículas cargadas una vez que se han producido las colisiones. Esto permite determinar el momento a partir de la curvatura y el signo de la carga a partir del sentido de giro. En el caso de CMS generan un campo de 3.8T.
- **Renovación del haz:** Se emplean imanes para conducir los haces fuera del LHC y reducir su intensidad cuando se quieren renovar los protones que forman parte de ellos.

El LHC finalizó su construcción en el año 2008, momento en el que se iniciaron las primeras pruebas de funcionamiento. Un accidente ocurrido con uno de los sistemas eléctricos retrasó el inicio de las medidas físicas hasta finales de 2009. A partir de entonces la actividad del acelerador se ha dividido en periodos de funcionamiento conocidos como *Run* y periodos largos de parada conocidos como *Long Shutdowns* (LS), tal y como se representa en la figura 2.2. Las etapas de *Run* no consisten en tomas de datos constantes, ya que buena parte de las semanas se emplean en la puesta a punto y el ajuste del hardware, detectores y haces de partículas, así como paradas técnicas para solucionar problemas. Durante estos periodos la mayor parte de las medidas se realizan con colisiones de protones, reservando unas pocas semanas a final de año para los iones. Los *Runs* se separan con paradas de larga duración, alrededor de 2 o 3 años, en los que se realizan mejoras en todos los subsistemas y se busca aumentar la energía en centro de masas [9].

En la actualidad el LHC se encuentra en el *Run* 3, iniciado a mediados de 2022 y cuya finalización se estima alrededor de finales del año 2025.

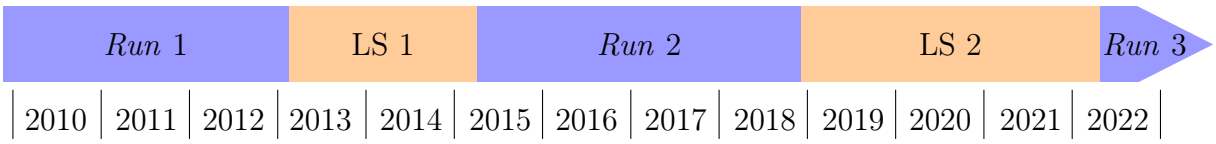


Figura 2.2: Periodos de funcionamiento del LHC hasta la actualidad. Fuente: Elaboración propia

Los datos empleados en este trabajo se han recogido durante el *Run* 2, en concreto entre los meses de mayo y octubre de 2018. Esto ha permitido que los datos se encuentren validados y clasificados por expertos según la calidad de los mismos de cara a realizar análisis físicos, como veremos más adelante.

Sin embargo, el objetivo de este trabajo no se limita al *Run* 2 ni a realizar un análisis retrospectivo sobre los datos recogidos previamente, sino que se busca estudiar la validez de un método que puede ser aplicado en los sucesivos *Runs* con el fin de mejorar la eficiencia y el acierto en el proceso de control de los datos del LHC.

Como se ve en la figura 2.3, la planificación actual del LHC incluye su funcionamiento hasta el *Run* 6, extendiéndose al menos hasta 2041. Esta incluye a más corto plazo un plan de mejora del acelerador para aumentar su luminosidad, es decir, el número de colisiones posibles por unidad de tiempo. Este proyecto, conocido como LHC de alta luminosidad (HL-LHC) podría finalizar su desarrollo alrededor de 2029 [10].

Este plan se centra en la mejora de numerosos componentes, tales como los tripletes de imanes cuadrupolares de los detectores, los sistemas de enfriamiento de los electroimanes o aquellos encargados de la colimación de los haces. Este aumento en la luminosidad permitiría observar un mayor número de eventos menos frecuentes, como pueden ser los modos de producción y desintegración menos probables del bosón de Higgs.

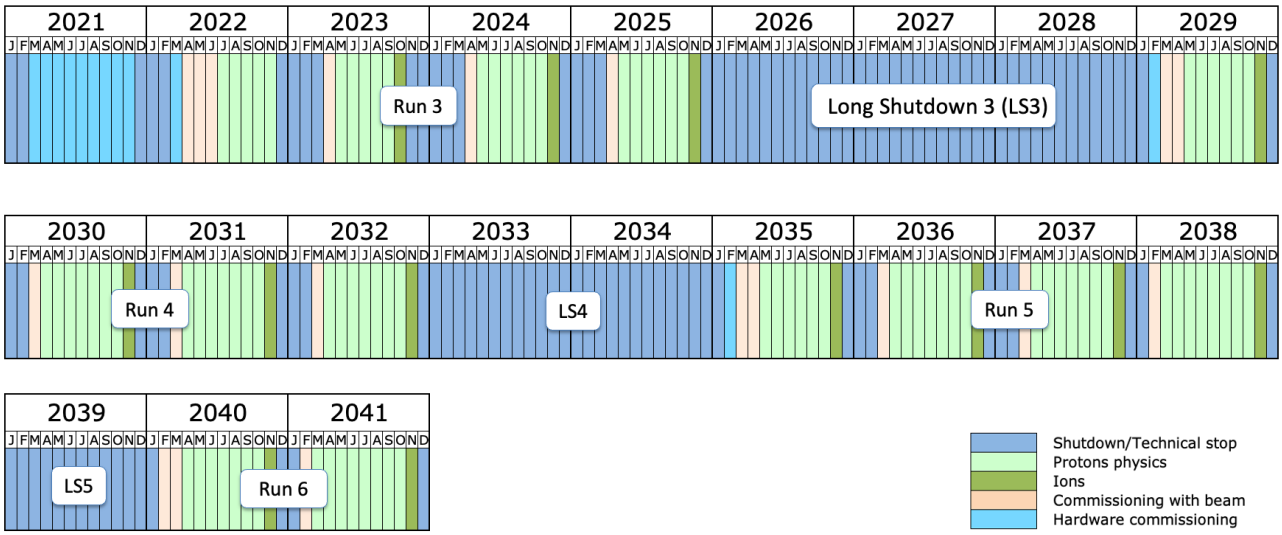


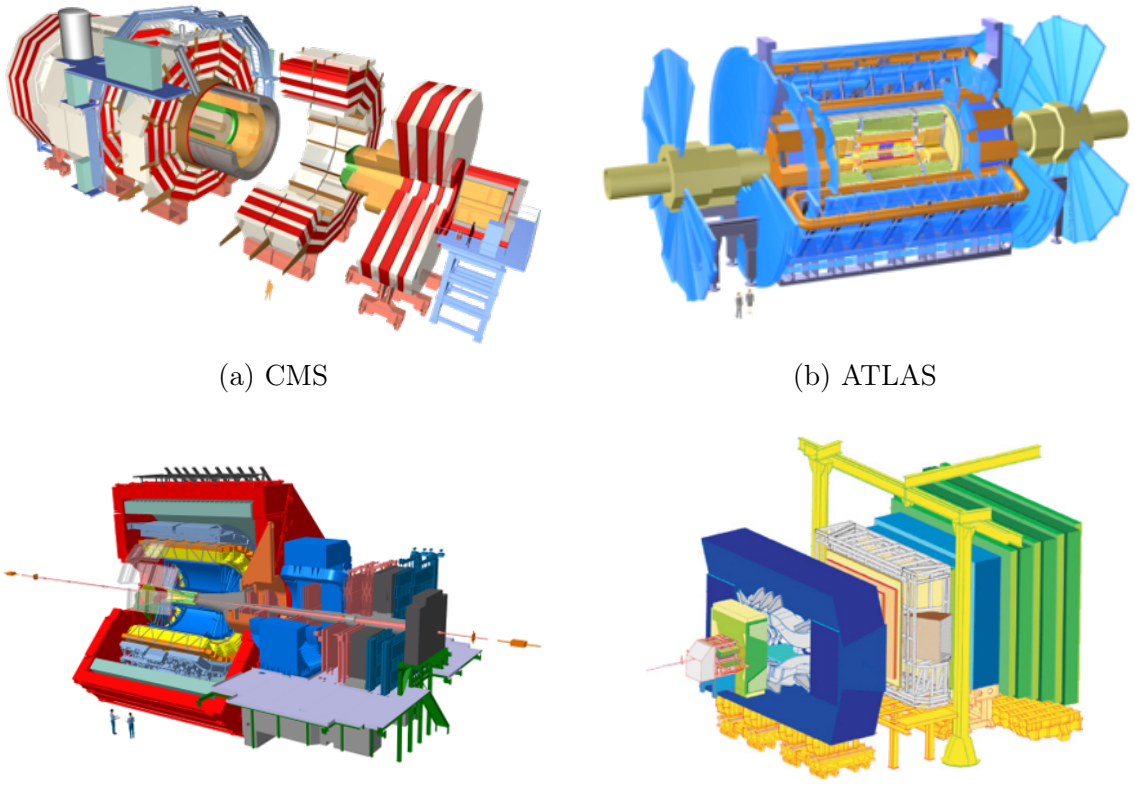
Figura 2.3: Planificación a largo plazo del funcionamiento de LHC. Fuente: [11]

2.2. Experimentos en el LHC

En total existen nueve experimentos en el LHC. Cuatro de ellos (CMS, ATLAS, ALICE y LHCb) son los de mayor tamaño y poseen sus propias galerías excavadas alrededor de los cuatro puntos de colisión. Existen también cinco de menor tamaño (LHCf, TOTEM, MoEDAL, FASER y SND@LHC) que se encuentran situados cerca de los anteriores para aprovechar las colisiones producidas en ellos.

- **CMS (Compact Muon Solenoid):** Es un detector de propósito general, de manera que busca cubrir un amplio espectro de experimentos de física, incluyendo la búsqueda de física más allá del modelo estándar o el estudio del bosón de Higgs. Su aspecto más significativo es el solenoide de gran tamaño que posee, permitiendo generar un potente campo magnético en su interior. Se caracteriza también por ser un detector muy compacto, teniendo un tamaño considerablemente reducido para su peso.

- **ATLAS (A Toroidal LHC ApparatuS):** Al igual que CMS, es un detector de propósito general, diferenciándose ambos en su diseño y características técnicas pero con los mismos objetivos científicos. En el caso de ATLAS el campo magnético se genera mediante un sistema de forma toroidal alrededor de la tubería central. Es el detector de mayor tamaño que se ha construido hasta la fecha.
- **ALICE (A Large Ion Collider Experiment):** Está especializado en analizar colisiones de iones de plomo, en las cuales se produce un plasma de quarks y gluones. Este es un estado de la materia en el cual se pueden observar quarks aislados, algo que no es posible de forma natural y que imita las condiciones del universo primigenio.
- **LHCb (Large Hadron Collider beauty):** Se encarga del estudio de la asimetría entre materia y antimateria a través del quark b y su antipartícula. El estudio se realiza mediante una serie de subdetectores distribuidos a lo largo de una longitud de 20 metros.



(a) CMS (b) ATLAS

(c) ALICE (d) LHCb

Figura 2.4: Principales detectores del LHC. Fuente: [7]

- **LHCf (Large Hadron Collider forward):** Se centra en el estudio de los rayos cósmicos, obtenidos de forma artificial en el laboratorio a partir de aquellas partículas que se dispersan en la dirección del haz de protones tras las colisiones. Es por ello que se sitúa muy cerca de ATLAS. Está formado por dos pequeños subdetectores.
- **MoEDAL (Monopole and Exotics Detector At the LHC):** Busca partículas exóticas, como los monopolos magnéticos, que son propuestas teóricas que indicarían la existencia de física más allá del modelo estándar. Está formado por 400 detectores de trazas de plástico centellador situados alrededor del punto de colisión de LHCb.
- **TOTEM (TOTal, Elastic and diffractive cross-section Measurement):** Estudia el comportamiento y la posible estructura de los propios protones que colisionan en el LHC. Aprovecha las colisiones producidas en CMS, extendiéndose varios cientos de metros alrededor de él.
- **FASER (ForwArd Search ExpeRiment):** Ha comenzado a funcionar en el *Run* 3 y se centra en la búsqueda de partículas ligeras y que interactúan débilmente. Este tipo de partículas han sido propuestas por alguna teorías más allá del modelo estándar para explicar nociones como la materia oscura, la asimetría entre materia y antimateria o el origen de las masas de los neutrinos. Se encuentra junto a ATLAS.
- **SND@LHC (Scattering and Neutrino Detector at the LHC):** Al igual que FASER, ha comenzado a tomar medidas en el *Run* 3 y se localiza al lado de ATLAS. Busca estudiar los neutrinos producidos en las colisiones, unas partículas elementales sin carga y que apenas interaccionan con la materia, lo que ha hecho que aún no se hayan detectado de forma directa en el LHC.

2.3. El detector CMS

El *Compact Muon Solenoid* (CMS) es un detector de propósito general, cubriendo por tanto un amplio rango de experimentos, incluyendo algunos relacionados con el estudio del modelo estándar, la materia oscura o la búsqueda de dimensiones extra [12].

Su nombre resulta un buen resumen de algunas de las características de este detector:

- **Compacto:** Presenta una longitud de unos 28 metros y una altura de alrededor de 15 metros, mientras que su peso es de 14000 toneladas. Esto son unas dimensiones muy reducidas para ese peso, especialmente en comparación con otros detectores como ATLAS.

- **Muones:** Una de sus principales ventajas es la gran precisión que tiene para la detección de estas partículas elementales.
- **Solenoide:** Este dispositivo presente en CMS es el más grande de su tipo jamás construido, permitiendo crear un campo magnético de 3.8T en su interior.

Antes de iniciar su construcción, se plantearon cuáles eran algunos de los problemas en los que debía participar este detector, entre los que se incluyeron la búsqueda del bosón de Higgs, la física más allá del modelo estándar, la existencia de dimensiones extra o la búsqueda de partículas supersimétricas. Además de todo ello, debía ser capaz también de estudiar la física que tiene lugar en las colisiones de iones pesados [13, 14].

Todos estos hechos llevaron a imponer una serie de requisitos técnicos que debía cumplir el detector para poder llevar a cabo esas tareas:

- Muy buena capacidad para identificar muones. Alta resolución para la determinación de su momento y una correcta identificación del signo de su carga para p $<$ 1 TeV/c.
- Alta eficiencia en la reconstrucción de las trayectorias de las partículas producidas.
- Buena resolución para el estudio de la energía electromagnética y de la energía transversal faltante (utilizada para determinar indirectamente los neutrinos producidos).
- Cobertura del mayor ángulo polar posible, al menos hasta $|\eta| < 2.5$.

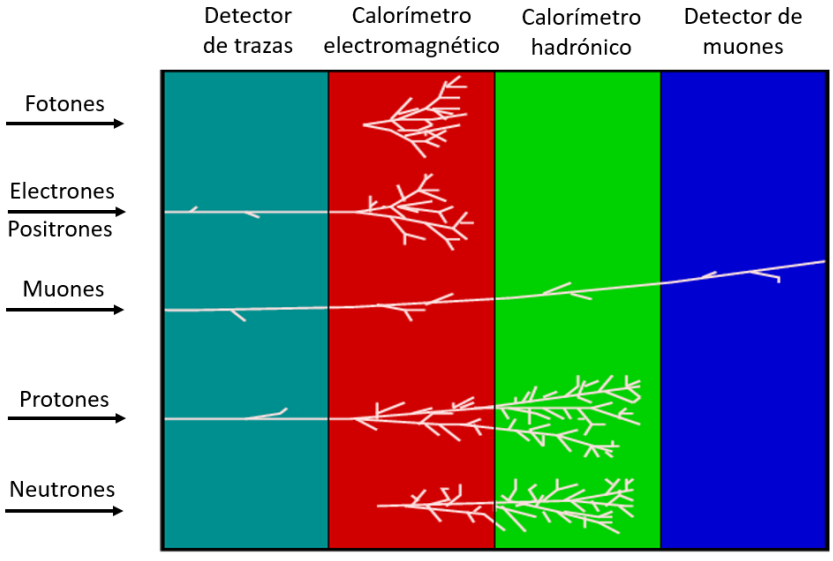


Figura 2.5: Interacción de partículas con un detector genérico de altas energías. Fuente: [15]

Con el fin de cumplir con los objetivos anteriores, CMS tiene una serie de capas concéntricas, de manera que la identificación y determinación de las propiedades de las partículas se hace combinando la información recogida en cada una de ellas. Un ejemplo de ello se muestra en la figura 2.5, en la aparecen las partículas que se medirían en cada capa de un detector genérico. Por su parte, la figura 2.6 muestra las capas que forman CMS.

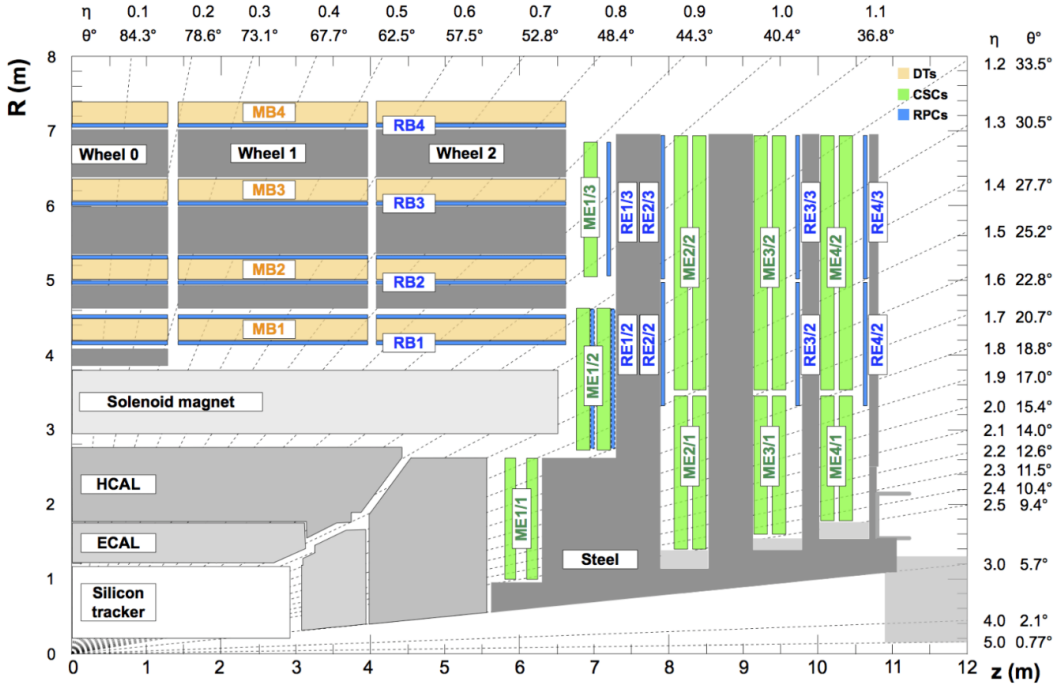


Figura 2.6: Corte longitudinal de un cuadrante de CMS. Fuente: [16]

Como ya se ha comentado previamente, uno de los elementos principales del detector es el solenoide superconductor que permite medir el momento de las partículas cargadas y el signo de su carga. Es el más potente del mundo, consiguiendo generar un campo de 3.8T en el interior de CMS y almacenando una energía de 2.7 GJ. Su gran tamaño hace que todas las capas de subdetectores se encuentren contenidas en el interior del solenoide a excepción de los detectores de muones. Estos últimos se encuentran intercalados con láminas de hierro que rodean todo el solenoide con el fin de confinar el campo magnético en el interior del detector.

La primera capa del detector que se encuentran las partículas producidas en la colisión es el detector de trazas, un sistema cuya finalidad es reconstruir sus trayectorias a partir de un conjunto discreto de puntos de detección. Debido a su cercanía con los puntos en los que se producen las colisiones, los materiales con los que se construye esta parte de CMS deben ser resistentes a valores altos de radiación.

El detector de trazas presenta varias capas, la primera formada por píxeles de silicio (en un radio entre 3 y 16 cm) y las siguientes construidas a partir de tiras de silicio, ambas representadas en la figura 2.7. La interacción de las partículas cargadas con estos pequeños detectores produce señales eléctricas (*hits*) que son posteriormente amplificadas, permitiendo reconstruir la trayectoria a partir de estas detecciones individuales. Estos sistemas determinan la posición con una precisión de unos 10 μm.

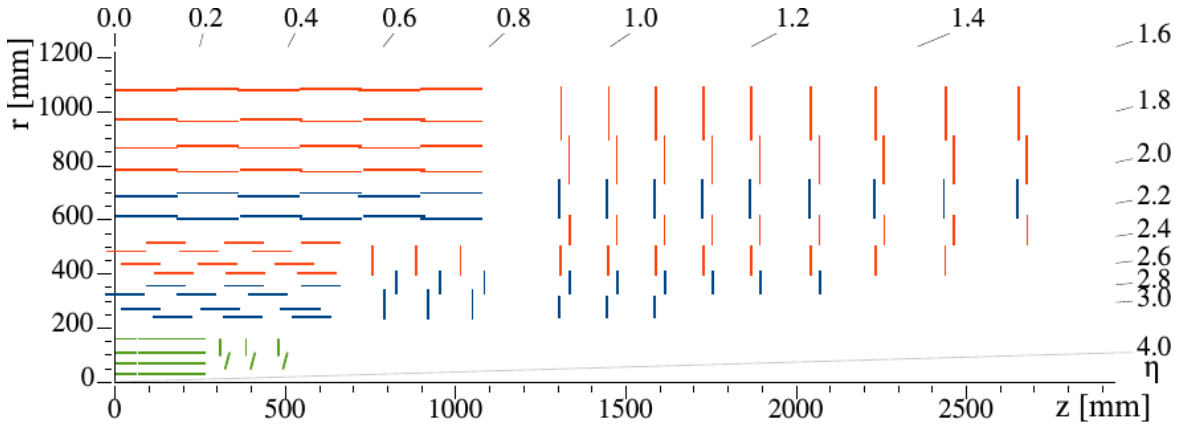


Figura 2.7: Corte longitudinal de un cuadrante del detector de trazas de CMS señalando los píxeles (verde) y tiras (rojo: una cara / azul: doble cara) de Si. Fuente: [17]

A continuación se sitúa el calorímetro electromagnético (ECAL), que permite la medición de la energía de electrones y fotones, y provoca que estas partículas se detengan en esta capa. Está formado por cerca de 70 000 cristales de $PbWO_4$, un material elegido porque permite obtener un calorímetro compacto, resistente a la radiación y con un tiempo de respuesta rápido. El funcionamiento se basa en el centelleo, es decir, que la interacción con electrones y fotones produce un destello cuya intensidad depende de la energía de las partículas. Estos destellos son recogidos por fotodetectores situados en cada uno de los cristales que los transforman en señales eléctricas.

La siguiente capa es el calorímetro hadrónico (HCAL), encargado de la detección de los hadrones (partículas formadas por dos o más quarks unidos por la interacción fuerte). Uno de sus requisitos es que debe ser hermético, impidiendo que las partículas producidas a partir de las desintegraciones de los hadrones y que no puedan ser detectadas en capas sucesivas escapen del calorímetro, con el fin de medir la energía faltante con precisión. Está organizado en capas alternas, unas de ellas fabricadas con material denso y absorbente, fundamentalmente latón, y otras consistentes en láminas de plástico centellador. La interacción de los hadrones con el material denso provoca la aparición de cascadas de partículas que generan la emisión de luz por el material centellador y que posteriormente será recogida por fibra óptica para su detección.

Como habíamos comentado previamente, una de las características principales de CMS es su capacidad para la identificación y el estudio de los muones. Estas partículas apenas pierden energía al pasar por las capas descritas anteriormente, por lo que atraviesan sin problema los calorímetros, siendo necesario incluir componentes específicos para identificarlos en la parte externa del detector. Cada muon es detectado a partir de las señales que genera en cada una de las cuatro capas que hay, que se encuentran separadas entre sí por las placas de hierro que confinan el campo magnético.

Existen un total de 1400 detectores, llamados cámaras de muones, y que se dividen en cuatro tipos: tubos de deriva (DT), cámaras de tiras catódicas (CSC), cámaras de placas resistivas (RPC) y multiplicadores gaseosos de electrones (GEM), siendo estos últimos una novedad para el *Run* 3. Su funcionamiento se basa en la ionización del gas que contienen por el paso de los muones, de manera que los electrones arrancados se recogen en un ánodo dando lugar a una corriente eléctrica. Sus diferentes características técnicas determinan su localización óptima en CMS.

2.4. El sistema de *trigger*

El enorme volumen de datos que se genera en las colisiones del LHC hace necesario llevar a cabo una reducción del mismo con el fin de permitir su recogida, almacenamiento y análisis. La importancia de este filtrado de datos no reside únicamente en facilitar su manejo, sino también en tratar de mantener únicamente aquellas colisiones que proporcionen información física interesante. El primer paso en este proceso es el sistema de *trigger*, diseñado para reducir este volumen de datos en al menos un factor 10^6 [18].

Este sistema consta de dos etapas: el *trigger* de nivel 1 (L1) y el *trigger* de alto nivel (HLT). El primero basa su funcionamiento en el uso de electrónica programable con el fin de hacer un filtrado basado en la búsqueda de sucesos con interés físico, por ejemplo debido a su alta energía. El segundo actúa a nivel de software combinando información de diferentes subdetectores de CMS para seleccionar los mejores eventos (resultados de las colisiones) que deben ser guardados.

Todos los eventos que no sean aceptados por el *trigger* no van a ser procesados y guardados, no existiendo la opción de recuperarlos. Este hecho remarca la importancia de que este sistema funcione correctamente para evitar la pérdida de información relevante.

El *trigger* L1 se divide en dos partes, una de ellas asociada a los calorímetros y otra a la detección de muones, dividiéndose a su vez cada una en otros tres niveles:

- **Nivel local:** Sistemas que reciben la información directamente de los componentes electrónicos individuales del detector. En el caso de muones se corresponde con cada una de las cámaras de muones y en los calorímetros con conjuntos apilados de subdetectores que reciben el nombre de torres.
- **Nivel regional:** Combina la información de varios componentes del nivel local. En el caso de muones permite reconstuir las trayectorias y en los calorímetros identificar las candidatas a ser partículas y sus energías.
- **Nivel global:** Recoge toda la información anterior para determinar los grupos de partículas (*jets*), la energía de los mismos y la energía faltante asociada a neutrinos.

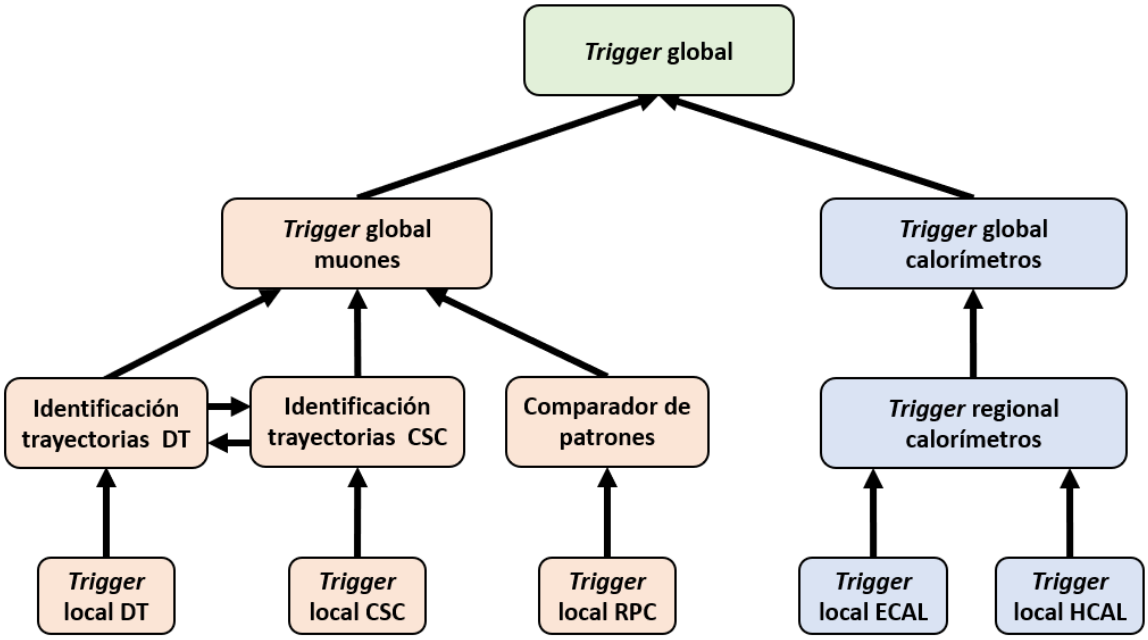


Figura 2.8: Estructura del *trigger* L1. Fuente: Elaboración propia.

Este nivel L1 obtiene datos con una frecuencia de 40 MHz, la frecuencia de trabajo del LHC, mientras que a la salida debe reducirla a un máximo de 100 kHz. Esto requiere un procesamiento extremadamente rápido de la información, obligando al uso de canales en paralelo y a tomar la decisión de aceptar o rechazar un suceso en unos pocos microsegundos [19].

La salida del L1 es transferida al HLT, que debe reducirla varios órdenes de magnitud hasta los 1000 Hz. Esto se realiza mediante algoritmos de software que permiten reconstruir los eventos a partir de las detecciones individuales. El criterio principal para el filtrado vuelve a ser el potencial del evento para aportar información física relevante. Los datos filtrados por el HLT son transferidos a un centro de computación para su posterior tratamiento *offline*.

2.5. *Data Quality Monitoring*

El grupo de monitorización de la calidad de datos (*Data Quality Monitoring*, DQM) tiene por objetivo, como su nombre indica, garantizar la calidad de los datos físicos recogidos [20, 21]. Para ello se encarga de tres tareas fundamentales:

- **Monitorización:** Se debe emitir una alarma cuando se detecta un mal funcionamiento de alguna parte del detector, ya sea a nivel de hardware o de software. Esta tarea debe realizarse a tiempo real.
- **Certificación:** Se analizan los datos obtenidos con el detector para determinar si la información contenida en ellos es válida o presenta algún tipo de error, clasificándolos como buenos o malos. No es necesario que esta tarea se realice a tiempo real.
- **Colaboración en la búsqueda de soluciones:** Se realizan informes acerca de las causas de estos problemas con el fin de facilitar que sean resueltos.

El trabajo del grupo de DQM comienza con una etapa de recolección de datos que se realiza en colaboración con los distintos grupos y expertos que trabajan en cada uno de los sectores de CMS. Estos datos, que están formados por eventos individuales, son tratados para obtener gráficas y medidas resumen que permitan realizar la monitorización y certificación. Estos resultados son procesados y se hacen disponibles a los grupos de trabajo de CMS a través de una interfaz que permite una consulta cómoda de los mismos, mantenerse actualizado sobre su evolución y establecer alarmas en caso de que los valores medidos no cumplan los requisitos mínimos.

El trabajo del DQM forma parte de todas las etapas del proceso de lectura de datos y reconstrucción de eventos a partir de ellos, siendo una pieza clave para garantizar que los datos a partir de los cuales se extraen las conclusiones físicas cumplen las condiciones para garantizar la calidad de estas.

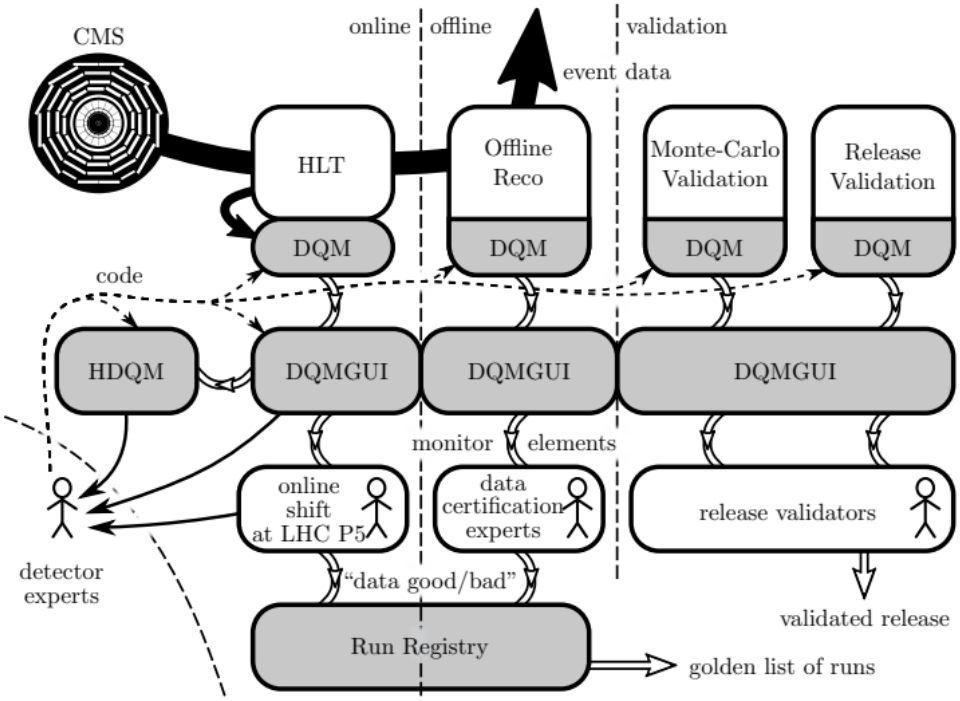


Figura 2.9: Resumen de la organización del DQM. Fuente: [20]

Como se puede apreciar en la figura 2.9, el DQM se encuentra presente desde el comienzo de la cadena de tratamiento de datos, en el *trigger* de alto nivel. La tarea en este punto se centra principalmente en la monitorización a tiempo real para garantizar que la reconstrucción se realiza de forma correcta. La enorme cantidad de información manejada en este punto hace que únicamente unos pocos histogramas puedan ser controlados de forma manual.

El DQM también se encarga de controlar la reconstrucción *offline* de eventos, las simulaciones realizadas mediante el método de Monte-Carlo y las nuevas versiones del código que controla el funcionamiento del detector (*release validation*) [22].

Los resultados de este trabajo se representan mediante histogramas y, en el caso de la certificación, mediante los resultados de aplicar test a dichos histogramas para determinar la calidad de los mismos. Estas representaciones se pueden consultar a través del DQMGUI [1], que constituye un primer almacenamiento de información y permite su consulta a través de una interfaz web que se puede ver en la figura 2.10.

[1]GUI (del inglés, *Graphical User Interface*) es una forma de presentar la información en un ordenador consistente en el uso de imágenes, iconos y menús en lugar de emplear únicamente texto.

Figura 2.10: Ejemplo de visualización en el DQMGUI. Fuente: [20]

La información también se almacena en el DQM histórico (HDQM) para tener una visión más general de la evolución a lo largo del tiempo. Este no suele tener gran relevancia de cara a la monitorización en tiempo real o la certificación, tareas para las cuales la información relevante es la correspondiente a las colisiones recientes. Sin embargo, las series temporales pueden resultar de gran valor para la búsqueda de soluciones a los problemas detectados en esas dos etapas. También dispone de una interfaz web para la consulta que se puede ver en la figura 2.11.

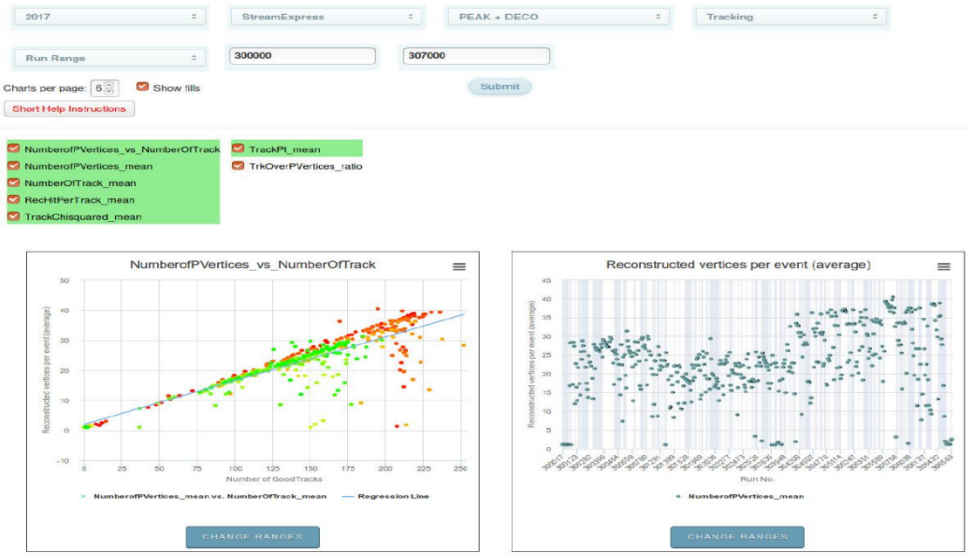


Figura 2.11: Ejemplo de visualización en el HDQM. Fuente: [23]

Finalmente, la última etapa consiste en el registro de *runs*, en el que se recoge la información proveniente de los distintos grupos del DQM con el fin de producir la "lista dorada" de *runs*, el conjunto de datos adecuados para el análisis físico.

El gran número de tareas en las que está involucrado el DQM pone de manifiesto la relevancia que tiene en el correcto funcionamiento de CMS. En concreto, este trabajo se centra en la certificación, el etiquetado de datos como buenos o malos. Esto se realiza fundamentalmente de forma manual mediante la revisión de los histogramas resumen por expertos. Un hecho relevante es la granularidad con la que se realiza dicha revisión, para lo cual necesitamos definir algunos conceptos:

- ***fill:*** Periodo que transcurre desde que se introduce un grupo de protones en el LHC hasta que se extrae porque la reducción en la luminosidad debida a las colisiones hace que no resulte interesante mantenerlos. Su duración es muy variable, pero en general es de varias horas, siendo esto decidido por los expertos del LHC.

- ***run:*** Intervalo de tiempo en el que CMS toma medidas en condiciones estables. Su duración también es muy variable, pudiendo ir desde unos pocos minutos hasta varias horas, siendo decidido esto por los expertos de CMS. Su finalización puede deberse a la detección de un error o cambios en los subdetectores o en el *trigger*, entre muchas otras causas. Para diferenciarlo del *Run* de LHC (tiempo entre dos paradas largas) se escribirá uno con su primera letra mayúscula y el otro con minúscula a lo largo del trabajo.

- ***lumisection:*** Fracción de un *run* en el que los protones realizan 2^{18} vueltas en LHC, teniendo una duración aproximada de 23 segundos. [2]

Los datos recogidos en CMS toman la *lumisection* como la unidad básica de identificación para agrupar varios eventos relacionados. Sin embargo, debido al enorme volumen de datos, la certificación se realiza por *runs*, de manera que los histogramas y etiquetas determinadas actualmente por expertos humanos van asociados a un *run* completo. Este hecho tomará especial relevancia en la propuesta del uso de aprendizaje automático en la certificación de datos y en las conclusiones de este trabajo.

[2]En la bibliografía es habitual que la palabra *lumisection* se abrevie como LS, al igual que se hace con *Long Shutdown*. Para evitar confusión, en este trabajo no se empleará dicha abreviatura.

Capítulo 3

Aprendizaje automático

3.1. Introducción al aprendizaje automático

Habitualmente, cuando se quiere resolver un problema con un ordenador se utiliza un algoritmo, consistente en una secuencia de pasos ordenados que se siguen para producir una información de salida a partir de la de entrada. Pensemos por ejemplo en la forma de escribir un programa que determine el máximo común divisor de dos números. Este estará formado por una serie de pasos bien conocidos que permitirá obtener el valor exacto de dicha operación una vez que se siguen.

Sin embargo, esto no es siempre posible para todos los problemas. Basta pensar por ejemplo en el reconocimiento facial, ya que no es posible determinar una serie de pasos que, implementados en un ordenador y aplicados de forma ordenada, permitan garantizar que el resultado es correcto. En casos como este nos podemos conformar con obtener aproximaciones basadas en analizar patrones que se puedan encontrar en los datos. Estos son los problemas en los que resulta especialmente relevante la aplicación del aprendizaje automático.

El aprendizaje automático (conocido habitualmente como *machine learning* o ML) consiste en la programación de un modelo matemático que dependa de ciertos parámetros que serán ajustados utilizando datos de entrenamiento, por ejemplo procedentes de la experiencia pasada. Se dice que el modelo es descriptivo si se limita a explicar y obtener información a partir de los datos que se le proporcionan y se dice que es predictivo si a partir de ellos realiza suposiciones sobre la evolución futura [24, 25].

Existen distintas clasificaciones de los métodos de aprendizaje automático, pero una de las más habituales incluye las siguientes categorías:

- **Aprendizaje supervisado:** En el conjunto de datos de entrenamiento se proporcionan las entradas y la salida que se debería proporcionar para dichas entradas.
- **Aprendizaje no supervisado:** El conjunto de datos de entrenamiento está formado únicamente por las entradas, de manera que debe ser el modelo el que encuentre patrones, por ejemplo, para agrupar los datos.
- **Aprendizaje reforzado:** La salida del sistema es un conjunto de acciones, de manera que no es importante cada acción individual, sino que el conjunto de ellas permitan maximizar la recompensa.

La separación entre el aprendizaje supervisado y no supervisado no siempre es exacta, ya que existen numerosas situaciones en las que hay particularidades relacionadas con las etiquetas proporcionadas al modelo, por ejemplo debido a que algunas de ellas pueden ser erróneas o solo se proporciona una parte. Esto da lugar a hablar en ocasiones del aprendizaje semisupervisado.

No es posible acotar la gran cantidad de problemas a los que se puede aplicar el aprendizaje automático. A pesar de ello, podemos clasificarlos en dos grupos en función de su salida: clasificación y regresión. Los problemas de clasificación son aquellos en los que el número de posibles resultados del modelo es finito, hablando de clasificación binaria cuando únicamente hay dos valores. Por su parte, un problema de regresión es aquel en el que la cantidad de resultados posibles forma un continuo.

En este trabajo se trata con la certificación de datos de CMS, que deben ser etiquetados como buenos o malos, por lo que nos encontramos ante un ejemplo de problema de clasificación binario. Un ejemplo de problema de regresión podría ser la predicción de la cotización en bolsa de una empresa el próximo viernes.

A la hora de trabajar con modelos de aprendizaje automático existen algunas limitaciones que se deben tener en cuenta [26], entre las que se pueden destacar:

- **Explicabilidad:** Es habitual que los modelos de aprendizaje automático se comporten como ”cajas negras” en las cuales se introducen unos datos y proporcionan una salida, sin que el proceso intermedio pueda ser comprendido por un humano, lo que puede llevar a la imposibilidad de detectar y corregir errores.

- **Sesgos:** Resulta de vital importancia que los datos empleados en el entrenamiento sean representativos para evitar que el modelo adquiera sesgos durante su aprendizaje.
- **Número de parámetros:** Un número muy bajo de parámetros puede provocar la imposibilidad del modelo para adaptarse a los datos, mientras que un número muy elevado lleva a un ajuste muy bueno a los datos de entrenamiento pero la imposibilidad para generalizar su comportamiento. Estos reciben el nombre de subajuste y sobreajuste respectivamente.

Figura 3.1: Ejemplos de subajuste y sobreajuste. Fuente: [27]

3.2. Entrenamiento y evaluación del modelo

Una vez que se ha construido un modelo de aprendizaje automático, el siguiente paso es el entrenamiento del mismo. Este paso consiste en proporcionarle un conjunto de datos de manera que a partir de él se puedan ajustar los parámetros de los que depende el modelo. La metodología empleada para ello depende del modelo concreto pero puede realizarse, por ejemplo, minimizando una métrica entre los datos experimentales y la predicción del modelo para los ejemplos de entrenamiento disponibles.

Una vez superada esta etapa, el paso siguiente será la evaluación. Para ello se va a utilizar un conjunto de datos distinto del empleado para el entrenamiento, de manera que se pueda comprobar qué tal es el desempeño del modelo. Es muy importante que no se utilicen para la evaluación los mismos datos a los que se ha recurrido para el entrenamiento, pues sus parámetros ya están optimizados para ellos y los resultados no serían realistas. Es habitual que se tome el conjunto total de datos y se divida en un 70 % para entrenamiento y un 30 % para evaluación, si bien se pueden modificar libremente estas cantidades.

Como hemos comentado previamente, en nuestro caso estamos ante un problema de clasificación binaria. En ese caso la evaluación es sencilla, pues consiste en comparar los resultados que predice el modelo con las etiquetas asignadas por los expertos. Como los datos se clasifican en buenos y malos (que asociaremos a positivos y negativos respectivamente) tenemos cuatro posibles resultados en la evaluación:

- **Verdadero positivo (TP, *true positive*):** Datos etiquetados como buenos que nuestro modelo clasifica como buenos.
- **Verdadero negativo (TN, *true negative*):** Datos etiquetados como malos que nuestro modelo clasifica como malos.
- **Falso positivo (FP, *false positive*):** Datos etiquetados como malos que nuestro modelo clasifica como buenos.
- **Falso negativo (FN, *false negative*):** Datos etiquetados como buenos que nuestro modelo clasifica como malos.

Todos ellos se pueden representar mediante la matriz de confusión, cuyo aspecto general se puede ver en el cuadro 3.1. Esta es una tabla en la que las veces que acierta el modelo se recogen en la diagonal y las que falla, fuera de ella. Es por ello que, idealmente, debe tender a ser una matriz diagonal, siempre y cuando las etiquetas asignadas a los datos sean correctas.

		Predicción modelo	
		Bueno	**Malo**
Etiqueta	**Bueno**	TP	FN
	Malo	FP	TN

Cuadro 3.1: Matriz de confusión

A partir de los cuatro valores que aparecen en la matriz de confusión se pueden construir una serie de indicadores que permiten evaluar el modelo y comparar resultados, actuando a modo de medidas resumen [28].

- **Valor predictivo positivo (PPV, *positive predictive value*):** Permite conocer la tasa de datos correctamente etiquetados como buenos respecto al total de datos etiquetados como buenos. Interesa que este valor sea lo mayor posible. En ocasiones también se le conoce como precisión.

$$PPV = \frac{TP}{TP + FP}$$

- **Fracción de verdaderos positivos (TPR, *true positive rate*):** Expresa la tasa de datos buenos correctamente etiquetados respecto al total de datos buenos que existen. Nos interesa maximizar esta cantidad. También recibe el nombre de sensibilidad.

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

- **Fracción de verdaderos negativos (TNR, *true negative rate*):** Expresa la tasa de datos malos correctamente etiquetados respecto al total de datos malos que existen. Nos interesa maximizar esta cantidad. Se conoce también como especificidad.

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP}$$

A partir de las tres medidas anteriores se pueden construir otras:

- **Valor F:** Relaciona el valor predictivo positivo y la fracción de verdaderos positivos en un único valor. Depende de un parámetro β que puede tomar cualquier valor positivo. Nos interesa que F sea lo más grande posible.

$$F_\beta = (1 + \beta^2)\frac{PPV \cdot TPR}{\beta^2 PPV + TPR}$$

  El valor β permite modificar la importancia dada a la precisión y la sensibilidad. En concreto, indica cuántas veces más importante se considera la sensibilidad respecto a la precisión. Los valores más habituales son $F_{1/2}$, F_1 y F_2.

- **Media geométrica (GM)**: Relaciona las tasas de verdaderos positivos y negativos a través de la raíz cuadrada de su producto. Nos interesa que sea grande.

$$GM = \sqrt{TPR \cdot TNR}$$

  Esta última suele resultar más adecuada para describir conjuntos de datos no equilibrados.

3.3. Aprendizaje automático en la certificación de datos

Los objetivos de la física de altas energías requieren la identificación de sucesos muy poco frecuentes entre una cantidad muy elevada de datos proporcionados por los detectores en los aceleradores de partículas. Esta cantidad de información producida crece con el avance de las tecnologías de detección y su estudio con técnicas tradicionales de análisis resulta cada vez más complejo. Es por ello que, desde hace ya un tiempo, la aplicación de modelos de aprendizaje automático forma parte del trabajo diario [29].

Entre estas tareas en las que las técnicas de ML comienzan a tener un papel importante se encuentran las simulaciones realizadas para comparar los resultados experimentales con las predicciones del modelo estándar, ya que debido a su elevado coste computacional requieren del uso de métodos aproximados. Los modelos de aprendizaje automático buscan aumentar la precisión de los mismos manteniendo un tiempo razonable de simulación. Otras áreas serían la mejora del tiempo de actuación del *trigger*, la detección de errores en detectores y sistemas informáticos o la reconstrucción de partículas y sus propiedades a partir de los productos de su desintegración, donde las redes neuronales han tomado el relevo a los árboles de decisión [30].

En el caso de la certificación de datos existen dos razones principales que invitan a la implementación de métodos de aprendizaje automático [31]:

- La gran cantidad de datos generados obliga a los responsables del DQM a revisar manualmente cientos de histogramas, lo que puede llevar a generar fatiga en los trabajadores, resultando en errores humanos en el etiquetado.
- La certificación se realiza *run* a *run*, lo que supone una pérdida de granularidad respecto a un posible etiquetado de cada *lumisection*. Esto lleva a que se puedan encontrar *lumisections* malas en *runs* marcados como buenos y *lumisections* buenas en *runs* etiquetados como malos.

Todo ello lleva a proponer técnicas de aprendizaje automático que permitirían obtener una mayor granularidad y aliviar la carga de trabajo de los responsables del DQM, además de aumentar la fiabilidad y el tiempo de respuesta de la certificación de datos. Algunas de las propuestas estudiadas en la actualidad incluyen el análisis de componentes principales (PCA), la factorización no negativa de matrices (NMF) [32] y *autoenconders* (AEs) [33].

Capítulo 4

Análisis de Componentes Principales

4.1. Introducción

El análisis de componentes principales (PCA por sus siglas en inglés) es una técnica que permite reducir la dimensionalidad de un conjunto de datos mediante la definición de nuevas variables de estudio, llamadas componentes principales (PC), de manera que se consigue facilitar la interpretación de los datos a la vez que se trata de minimizar la pérdida de información [34]. Para ello se apoya en la siguientes suposiciones:

- Las direcciones que contienen más información son aquellas que tienen mayor varianza.

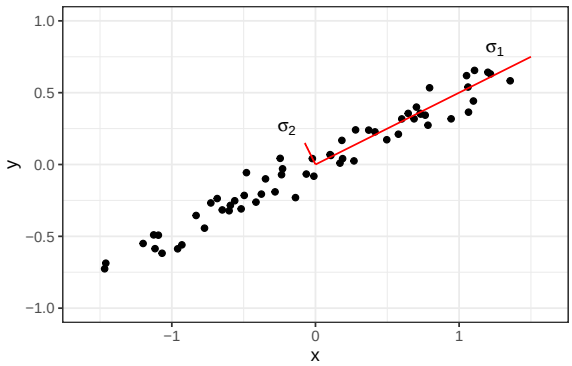


Figura 4.1: Importancia de la varianza en un gráfico bidimensional. Fuente: Elaboración propia

Supongamos que cada punto de la figura 4.1 es la posición de un objeto en un plano en cada instante de tiempo. Si queremos describir el movimiento quedándonos con una única dirección, la mejor será aquella a lo largo de la cual los datos están más dispersos, siendo esta la de mayor varianza (σ_1^2) frente a σ_2^2, que podría deberse por ejemplo a ruido.

- Las nuevas variables son combinación lineal de las anteriores.
- Las componentes principales son ortogonales entre sí.

Supongamos entonces que tenemos un conjunto de datos formado por n observaciones de m variables $X_1, ..., X_m$, algo que se puede representar mediante una matriz (o tabla) de n filas y m columnas en la que cada columna se corresponde con una de esas variables y cada fila es una observación, de manera que el elemento (i, j) se corresponde con la observación número i de la variable X_j [35, 36].

$$\begin{array}{c} \\ \text{Obs. } 1 \\ \text{Obs. } 2 \\ \vdots \\ \text{Obs. } n \end{array} \begin{array}{c} \begin{array}{ccccc} X_1 & X_2 & X_3 & \dots & X_m \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1m} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{nm} \end{pmatrix} \end{array} = X$$

Como hemos comentado, las m componentes principales Y_k $(k = 1, ..., m)$ se obtienen como combinación lineal de las variables iniciales:

$$Y_k = \sum_{j=1}^{m} \phi_{jk} X_j$$

Los coeficientes ϕ_{jk} son números reales y generalmente se toman para que sea una combinación lineal normalizada (la suma de cuadrados igual a 1). El objetivo es encontrar aquellas combinaciones que tengan una varianza máxima y que sean ortogonales entre sí.

Para realizar el cálculo se emplea la noción de covarianza. Si tenemos dos conjuntos de datos con media 0, $\{x_1, ..., x_n\}$ e $\{y_1, ..., y_n\}$, su covarianza es:

$$\sigma^2_{XY} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i$$

Esta medida representa la relación entre los conjuntos de datos. Si es elevada las variables tienen una relación lineal, mientras que si es igual a 0 serán linealmente independientes. La covarianza de un conjunto de datos con él mismo es la varianza. En nuestro caso, si las variables X_i tienen media 0, la matriz de covarianza se calcula como:

$$C = \frac{1}{n} X^t X$$

Esta es una matriz de dimensión m en la cual el elemento (p, q) va a ser la covarianza de las observaciones de las variables X_p y X_q. Por tanto, la diagonal estará formada por las varianzas de las medidas de las distintas variables.

Idealmente nos interesaría definir unas variables nuevas de manera que sus varianzas sean altas y sus covarianzas lo más bajas posibles. Esto permite que cada variable explique la mayor cantidad de información posible y a la vez no sean redundantes (si están relacionadas linealmente, habrá información repetida). Esto se corresponde con la diagonalización de la matriz de covarianza y los autovectores serán las componentes principales del conjunto de datos. El procedimiento se puede resumir de la siguiente forma:

1. Se crea una matriz donde cada columna es una variable y cada fila una observación.
2. Restar a cada columna su propia media, de manera que pasen a tener media 0.
3. Calcular la matriz de covarianza y diagonalizarla. Los autovectores serán las componentes principales y los autovalores sus varianzas.

De esta forma, para reducir la dimensionalidad del conjunto de datos (es decir, el número de variables con las que estamos trabajando) basta quedarnos con aquellas componentes principales que tengan mayor varianza, con lo que nos estaremos asegurando de reducir al mínimo la pérdida de información, al restringirnos a aquellas que son más descriptivas de los datos.

Un aspecto muy interesante del análisis de componentes principales es que permite reconstruir los datos una vez que se han eliminado las componentes principales menos descriptivas. Esto permite obtener una aproximación de los datos en un espacio de dimensión menor, lo que equivale a proyectarlos en dicho espacio. Recuperando el ejemplo anterior:

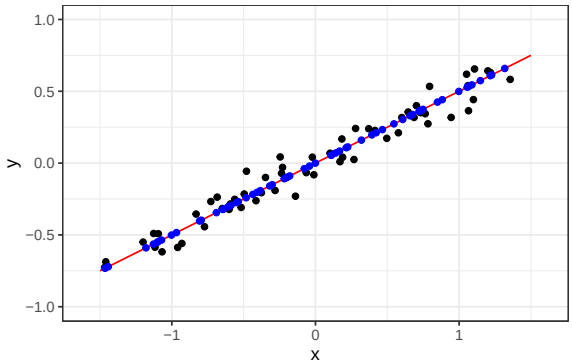


Figura 4.2: Reconstrucción de puntos (azul) con una única PC. Fuente: Elaboración propia

En la figura 4.2 se puede ver un ejemplo de la reconstrucción de los puntos del primer ejemplo. Estos estaban inicialmente descritos por dos variables x e y y su primera componente principal es $(2/\sqrt{5})x - (1/\sqrt{5})y$, una combinación lineal de ambas. La aproximación de dichos puntos con una única componente principal reduce el espacio de dos dimensiones a una, pudiendo obtener una aproximación de los puntos iniciales mediante su proyección, en color azul.

Matemáticamente, la reconstrucción de los datos X usando r componentes principales se calcula como:

$$\widetilde{X} = XAA^t$$

donde $\widetilde{X}$ es la matriz con los datos reconstruidos y A es una matriz de m filas y r columnas donde cada columna contiene los coeficientes ϕ_{jk} asociados a una de las r componentes principales con las que nos hemos quedado. En el caso de que se haya restado la media a los datos iniciales para conseguir que tuvieran media 0, esta se tiene que volver a sumar en este momento.

4.2. Estructura de los datos

El objetivo a partir de este punto es aplicar el análisis de componentes principales para diseñar un sistema basado en aprendizaje automático orientado a la certificación de los datos de los muones registrados en CMS. Con el fin de diseñar y comprobar el comportamiento del modelo se va a disponer de un conjunto de datos que vamos a describir en esta sección.

El conjunto de datos empleado en este trabajo consta de cuatro subconjuntos llamados eras y que están etiquetados mediante las letras A, B, C y D. La estructura de la información es la misma en todos ellos y la diferencia es que son mediciones realizadas en distintos periodos de funcionamiento del detector entre mayo y octubre de 2018, un periodo que se corresponde con el *Run* 2 del LHC. Cada una de las eras contiene varios *runs* de CMS y las correspondientes *lumisections* dentro de cada uno de ellos.

Debido a que los *runs* son consecutivos dentro de cada era, podemos considerar que estamos trabajando con una serie temporal, de manera que se puede estudiar la evolución de unos *runs* a otros. **Con el fin de facilitar el manejo de los datos se han renombrado los *runs* de menor a mayor** comenzando en 1, si bien se ha conservado la relación entre los identificadores numéricos originales y los nuevos para poder revertir el cambio si fuera necesario.

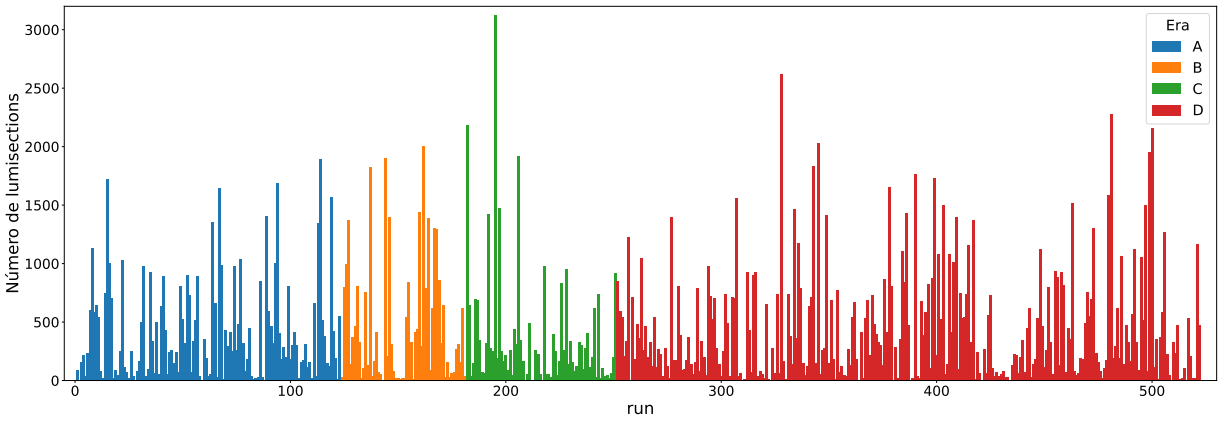


Figura 4.3: Número de *lumisections* en cada *run* para cada era.

Se puede observar claramente en la figura 4.3 que no todos los *runs* tienen el mismo número de *lumisections*, incluyendo algunos con menos de 10 hasta otros con varios miles de ellas. La era D es la más numerosa, siendo la A la siguiente con mayor número de *lumisections*.

Los datos incluyen medidas correspondientes a diversas propiedades de los muones detectados en CMS: momento total, momento transverso, ángulo azimutal, pseudorapidez y coeficiente χ^2/g.l. . La primera de ellas no será utilizada, puesto que a nivel de certificación nos basta con incluir uno de los dos momentos. El coeficiente χ^2/g.l. proviene de la reconstrucción de trayectorias en el detector de trazas, pues es una magnitud que mide el grado de ajuste entre los puntos experimentales y la trayectoria reconstruida. Como es habitual en certificación, para analizar toda esta información se trabajará con histogramas.

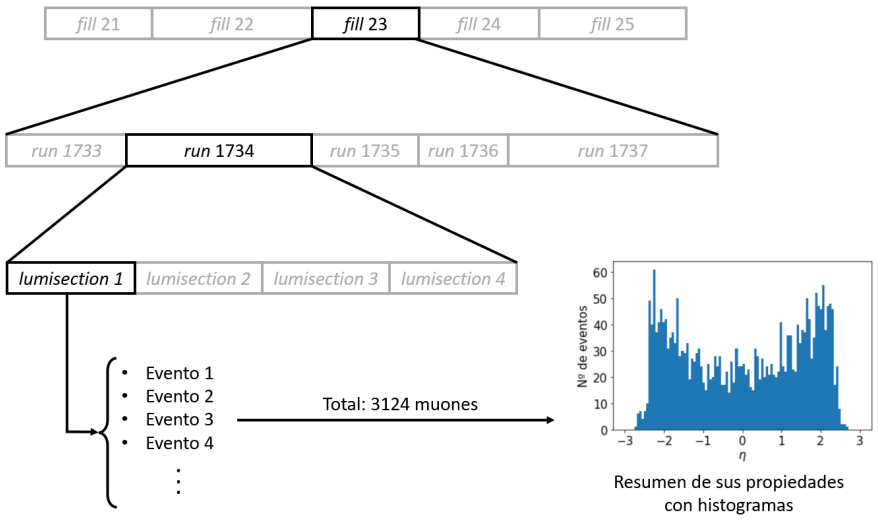


Figura 4.4: Resumen de la estructura de los datos. Fuente: Elaboración propia

La figura 4.4 es un resumen de los datos de muones con los que estamos trabajando[1]. En cada *fill* de LHC existen varios *runs* de CMS en los que el detector toma datos de forma estable. Como se comentó previamente, estos se dividen a su vez en periodos de unos 23 segundos llamados *lumisections*. Cada una de ellas estará formada por un número de eventos que depende de las colisiones que hayan tenido lugar, interesándonos en nuestro caso el número de muones que se hayan producido y resumiendo cada propiedad de los mismos mediante histogramas.

Uno de los aspectos fundamentales es la clasificación de los datos en buenos y malos, una separación que se realiza en base a la similitud de las medidas con unos estándares de calidad. Existen diversos motivos por los cuales unos datos pueden ser clasificados como malos, pero en numerosas ocasiones se debe a un mal funcionamiento de alguno de los sistemas de recolección de datos, ya sea a nivel del detector, *trigger* o software.

El conjunto de datos presenta cada *run* etiquetado como bueno o malo, una clasificación que ha sido realizada manualmente por expertos a partir del análisis de los correspondientes histogramas. Serán estas etiquetas las que nos permitirán el entrenamiento y evaluación de nuestro modelo en un contexto de aprendizaje semisupervisado, como veremos más adelante.

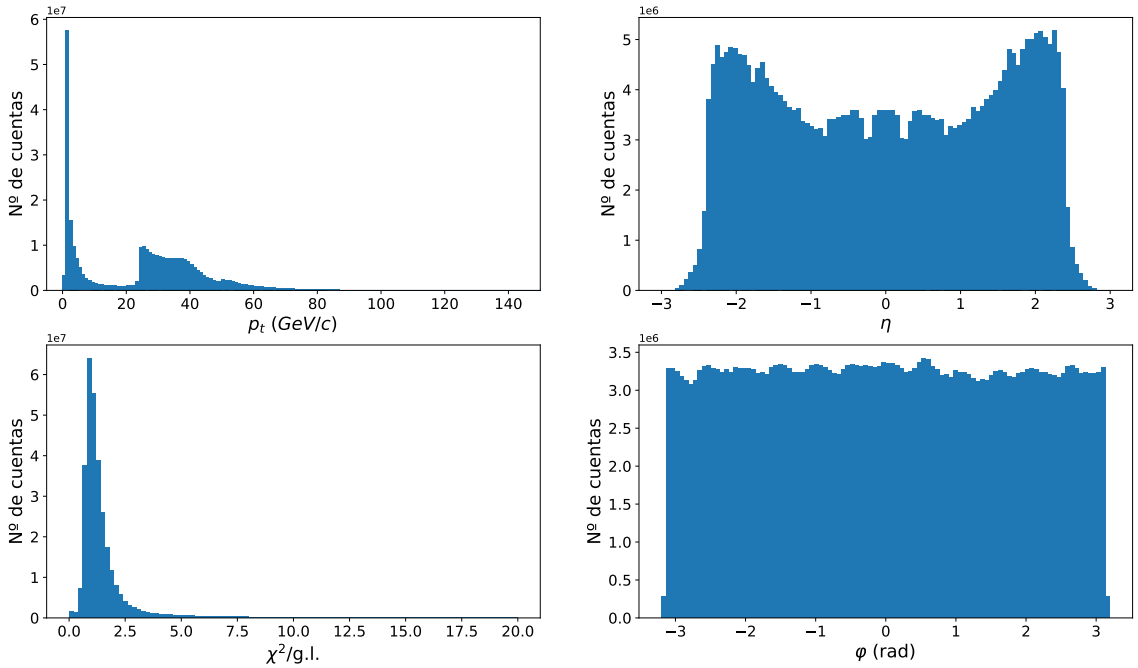


Figura 4.5: Histogramas para los *runs* etiquetados como buenos en la era A.

[1]Los números asignados a los fills y *runs* en la figura son una licencia artística elegida por simplicidad. Los valores actuales de ambos identificadores son mucho más altos.

Como se ha comentado, la certificación de datos se realiza mediante la comparación de los histogramas obtenidos con aquellos que se deben tener en caso de un funcionamiento óptimo de todos los sistemas. Con el fin de hacerse a la idea del aspecto que deben tener los histogramas de las diferentes propiedades estudiadas, se representan en la figura 4.5 los correspondientes a los datos de todos los *runs* etiquetados como buenos en la era A.

La distribución del momento transverso p_t presenta dos regiones principales. Una de ellas se sitúa a la derecha y se caracteriza por comenzar en la región de los 25 GeV/c a causa del sistema de *trigger*, que impone una cota inferior para el momento que deben tener los muones para ser detectados. Esta zona presenta una tendencia decreciente del número de detecciones para p_t altos, aunque presenta una cola muy alargada formada por unas pocas partículas que alcanzan varios cientos de GeV/c. Es interesante notar también que esta región situada a la derecha presenta algunos pequeños máximos locales en torno a 40 y 50 GeV/c debidos a que la ventana en la cual el *trigger* recoge datos de los muones no es constante a lo largo de la toma de datos, sino que se va desplazando hacia la izquierda a medida que la luminosidad en el LHC decrece, aceptando muones de menor momento para maximizar la cantidad de colisiones recogidas.

La otra región se sitúa a la izquierda y se encuentra separada de la anterior por un "valle". Esta se caracteriza por tener muones de muy bajo momento que han sido detectados debido a que forman parte de eventos en los que se ha producido más de un muon, de manera que la cota inferior impuesta por el *trigger* ha sido superada gracias a aquellos muones de mayor momento que forman parte de dicho evento, llevando a detectar todos los muones involucrados en él.

La distribución del ángulo azimutal φ es aproximadamente constante en todo su rango entre $-\pi$ y π radianes. Esto es debido a que no existe ningún motivo que haga que unas direcciones sean preferenciales respecto a otras para los muones producidos tras la colisión. Las pequeñas variaciones que se observan son debidas fundamentalmente a fluctuaciones estadísticas y el diseño de CMS, que puede incluir algunas zonas con una eficiencia ligeramente menor de detección.

En el caso de χ^2/g.l., los datos proporcionados se corresponden con el cociente del estadístico χ^2 de ajuste entre su número de grados de libertad (g.l.). En ese caso se obtiene una distribución típica según la teoría estadística, en la cual se alcanza un máximo en 1 y existe una leve asimetría con cola hacia la derecha [37].

Si nos centramos ahora en la pseudorapidez η, se observa un total de cinco máximos. Los tres centrales, de menor tamaño, son debidos a la aparición de varios mínimos en la distribución que se explican mediante la geometría del detector. Como se ve en las figuras del capítulo 2, CMS presenta un total de 5 ruedas, que son secciones del detector que tienen un hueco entre ellas. Estas se encuentran en la zona central (*barrel*) del detector, donde están las cámaras de muones de tipo DT, y provoca que algunos muones puedan atravesar todas las capas de CMS sin ser detectados suficientes veces como para identificarlos adecuadamente.

Los dos máximos externos, que están situados alrededor de $\eta = \pm 2$ son causados por los muones de bajo momento, los cuales presentan una trayectoria con mucha curvatura por la acción del campo magnético. Así, aquellos que presentan una pseudorapidez baja (salen muy perpendiculares al haz de protones) no son capaces de alcanzar las cámaras de muones de la capa exterior debido a esa curvatura, mientras que los que tienen mayor pseudorapidez pueden llegar a las cámaras de muones situadas en los tapones (*endcaps*) del detector, donde están las cámaras de tipo CSC, a pesar de tener bajo momento. Esto hace que las detecciones de muones de alta pseudorapidez y bajo momento se vean favorecidas, siendo de hecho estos últimos los más numerosos, como se ve en la figura 4.5.

Una vez que se ha visto el aspecto general de los histogramas etiquetados como buenos, nos podemos centrar en la limpieza de los datos. Existen numerosas *lumisections* que presentan un recuento muy bajo de eventos y algunas de ellas están incluso vacías. Esto puede suponer un problema a la hora de entrenar el modelo debido a que sus histogramas están enormemente afectados por el ruido o carecen de información. Algunas de ellas se encuentran en *runs* etiquetados como buenos y otros como malos, por lo que será necesario comprobar todas ellas.

La decisión del umbral para determinar cuándo se descarta una *lumisection* por datos insuficientes es arbitraria. En este trabajo se ha decidido que se descartarán aquellas que tengan un número de muones inferior a un 5 % del valor medio de todas las *lumisections*. La media en cada una es de 6245, de manera que **se tomará el corte en 300 muones**.

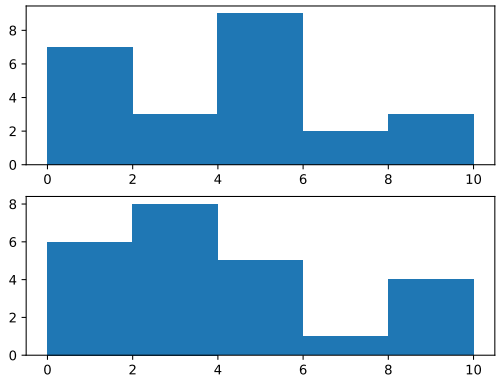
Era	A	B	C	D	Total
Datos malos (%)	0.48	0.76	1.48	0.47	0.62

Cuadro 4.1: Proporción de datos etiquetados como malos en cada era.

Un posible inconveniente en el entrenamiento y evaluación del modelo es la baja proporción de datos etiquetados como malos, como se recoge en el cuadro 4.1 para cada una de las eras y para el total. Esta asimetría entre datos buenos y malos podría provocar la aparición de sesgos en el aprendizaje del modelo, algo que se evitará trabajando en un contexto de aprendizaje semisupervisado en el que el modelo solo se entrenará con datos buenos.

4.3. Aplicación de PCA a la certificación

El primer paso para aplicar el análisis de componentes principales a los datos disponibles es construir la correspondiente matriz X. Vamos a considerar cada intervalo del histograma (habitualmente llamado *bin*) como una propiedad X_i, de manera que le asociaremos una columna de X y cada histograma será una fila de la matriz. Cada elemento de la matriz será simplemente el número de cuentas correspondiente a ese intervalo del histograma.



$$\longrightarrow \qquad X = \begin{pmatrix} 7 & 3 & 9 & 2 & 3 \\ 6 & 8 & 5 & 1 & 4 \end{pmatrix}$$

Figura 4.6: Ejemplo de transformación de dos histogramas en una matriz.

En el caso del conjunto de datos que se emplea en este trabajo, el número de columnas de la matriz será de 500 para p_t y de 100 para las otras tres propiedades, mientras que el número de filas es el número de *lumisections*, ya que tenemos un histograma para cada propiedad y para cada *lumisection*. El número de componentes principales utilizadas para resumir y aproximar los histogramas de cada propiedad puede variar entre 1 y el número de columnas.

Una vez obtenidas las componentes principales (PC) de X se pueden volver a reconstruir los histogramas pero únicamente utilizando el número de ellas deseado, de manera que serán una aproximación de los originales, siendo más precisa cuanto mayor sea el número de PC.

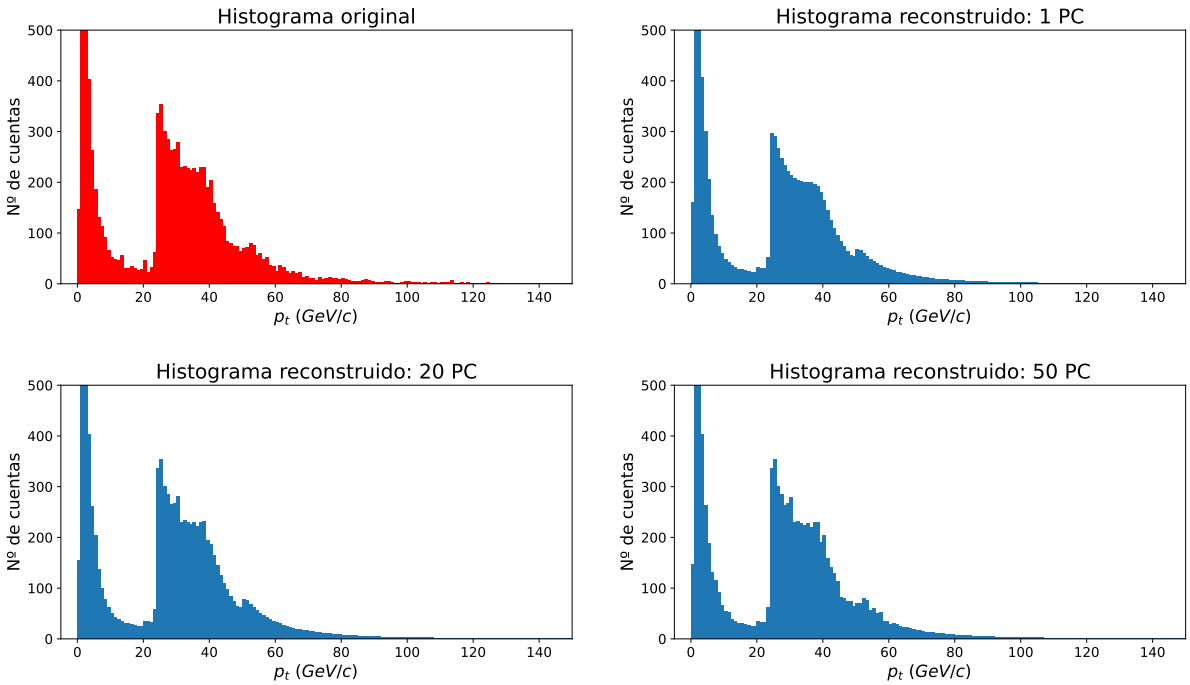


Figura 4.7: Reconstrucción según el número de componentes principales.

La figura 4.7 muestra perfectamente cómo una única componente principal ya recoge el comportamiento general del histograma en la reconstrucción, mostrando los picos más relevantes, si bien aparecen muy suavizados. Los detalles debidos a las particularidades de cada histograma se consiguen reconstruir cada vez mejor a medida que el número de componentes aumenta.

El funcionamiento de PCA para la detección de anomalías se basa en el hecho de que las componentes principales recogen la tendencia general de los datos, de manera que al realizar la reconstrucción eliminando algunas de ellas, el modelo va a ser incapaz de recuperar aquellos histogramas que se desvíen de esa tendencia.

Esto permite aplicar el análisis de componentes principales como un método de aprendizaje no supervisado, ya que será capaz de señalar aquellos histogramas que se desvíen del comportamiento mayoritario [38]. Sin embargo, esto solo garantiza éxito en la correcta clasificación si los datos en los que se entrena están muy desequilibrados, presentando un número muy superior de *lumisections* buenas. En nuestro caso podemos aprovechar que disponemos de las etiquetas de los expertos para **obtener la matriz A de componentes principales en el entrenamiento a partir de los datos buenos**, de manera que el modelo sea incapaz de reconstruir en la etapa de evaluación los histogramas que se desvíen del comportamiento de los buenos.

Para determinar la diferencia entre los histogramas originales y los aproximados será necesario definir una métrica. En nuestro caso se va a utilizar el error cuadrático medio (ECM):

$$E_i = \frac{1}{n}\sum_{j=1}^{n}(\widetilde{x}_{ij} - x_{ij})^2$$

Aquí E_i denota el ECM en la reconstrucción del histograma i-ésimo, siendo x_{ij} el valor que toma el histograma original en el intervalo j y $\widetilde{x}_{ij}$ su valor reconstruido. Así, los histogramas que se desvíen del comportamiento propio de los datos buenos tendrán un ECM alto.

El número de eventos que hay en cada histograma varía enormemente, habiendo algunos con 300 eventos tras la limpieza de datos, hasta otros con varios miles de muones detectados, lo que supone una dificultad añadida a la hora de comparar los resultados para las distintas *lumisections*, ya que aquellas que tengan un mayor número de eventos llevarán asociado habitualmente un ECM mayor pues una pequeña diferencia relativa en un intervalo lleva asociado un mayor error en términos absolutos. Con el fin de solucionar este inconveniente, **los histogramas se normalizarán a la hora de calcular el error de reconstrucción.**

Podemos ver como ejemplo la figura 4.8, en la que se representa el ECM para el momento transverso en las *lumisections* de la era B después de entrenar el modelo con los datos de la era C con tres componentes principales. En rojo se representan aquellas *lumisections* en los *runs* etiquetados como malos por los expertos. Usaremos este ejemplo para presentar el modelo.

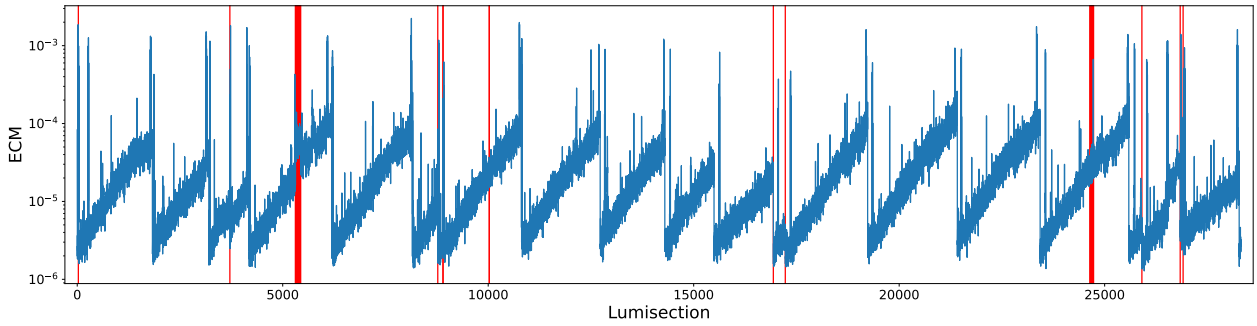


Figura 4.8: ECM por *lumisection* para el p_t en la era B. Entrenamiento con la era C.

Se puede ver una tendencia general en la que la mayor parte de las *lumisections* presentan un error acotado entre unos límites, aproximadamente entre 10^{-6} y 10^{-4}, mientras que algunas de ellas se salen de estos de forma significativa, generando una forma de ”agujas”.

Un hecho relevante es la estructura que aparece de forma repetida en la que el ECM aumenta de forma progresiva hasta llegar un punto en el que disminuye de forma brusca. Esto está relacionado con el funcionamiento del propio LHC, en el cual la luminosidad instantánea va reduciéndose durante su funcionamiento debido a las colisiones entre protones, haciendo que cada vez sean menos frecuentes. Este hecho explicaría el aumento en el error, ya que cuanto menor sea la luminosidad, mayor será el peso de la aleatoriedad, resultando en una mayor desviación de los histogramas del comportamiento ideal. Los saltos repentinos se deben a la introducción de los protones en el LHC en los *fills*, haciendo que la luminosidad aumente bruscamente.

Una posible forma de estudiar la luminosidad instantánea de forma indirecta con los datos disponibles consiste en determinar el número de muones que se han detectado en cada *lumisection*. Esto es una aproximación, ya que es una cantidad que depende de otros múltiples factores, tales como la eficiencia del detector o el tipo de colisiones que se hayan producido y su energía, pero aún así nos va a permitir observar claramente la correlación entre ambas magnitudes.

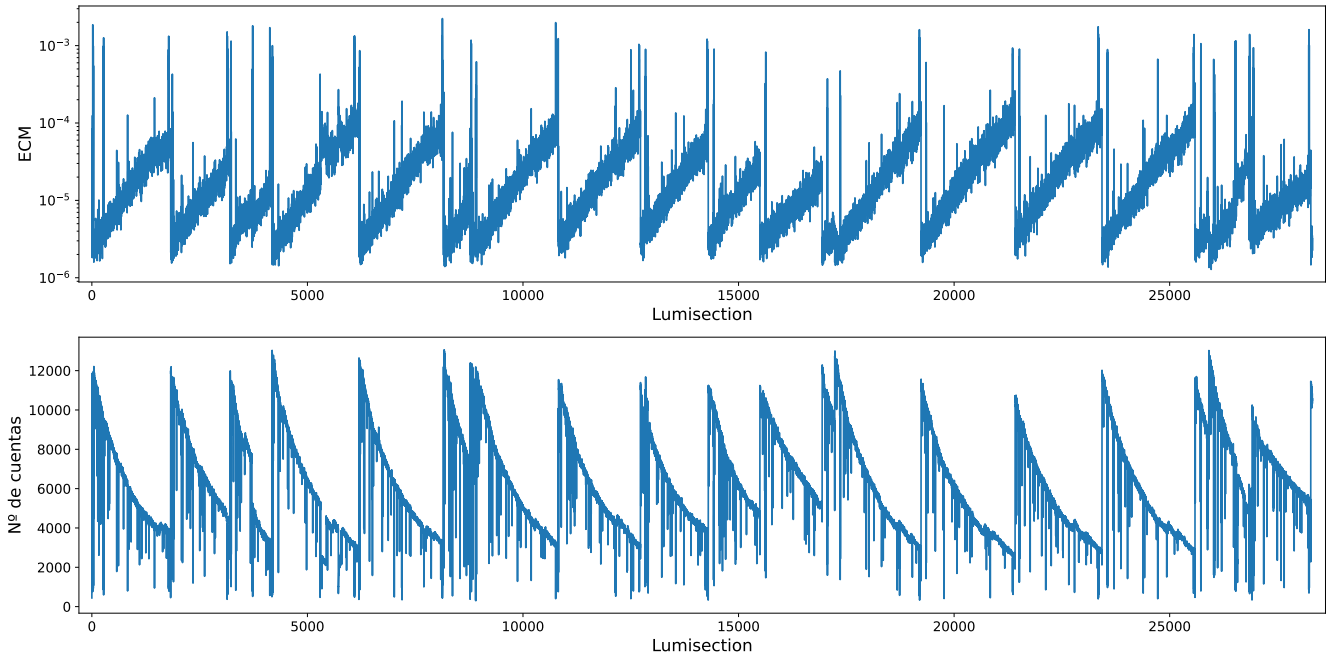


Figura 4.9: ECM para el p_t y número de muones por *lumisection* en la era B.

Vemos que la mayor parte de las desviaciones grandes en el valor del ECM (los picos que sobresalen significativamente) se encuentran en el paso de un *fill* a otro, es decir, en las zonas en las que se aprecia una discontinuidad en la luminosidad, probablemente debido a un funcionamiento no estable del detector durante este proceso.

En la gráfica inferior de la figura 4.9 se aprecia que la luminosidad tiene una tendencia general decreciente pero no lo hace de forma progresiva, sino que se aprecian caídas bruscas en ciertas *lumisections* que dan lugar a una estructura de picos dirigidos hacia la parte inferior. La explicación para este fenómeno son los intervalos de tiempo en los cuales no se pueden tomar medidas con la frecuencia necesaria para recoger todos los eventos, conocidos como tiempos muertos (*deadtimes*). Estos pueden estar causados por múltiples razones, como la saturación de la electrónica o fallos en los subdetectores.

En la figura 4.10 se representa en color negro la frecuencia de eventos aceptados por el *trigger* L1 y en color azul los *deadtimes*, pudiendo ver que la reducción en el número de detecciones está relacionada con esos tiempos muertos. Las líneas discontinuas amarillas se corresponden con el prescalado del *trigger*, es decir, el proceso en el que se relaja la exigencia sobre el momento mínimo que deben tener los muones para ser aceptados y que tiene lugar a medida que se reduce la luminosidad.

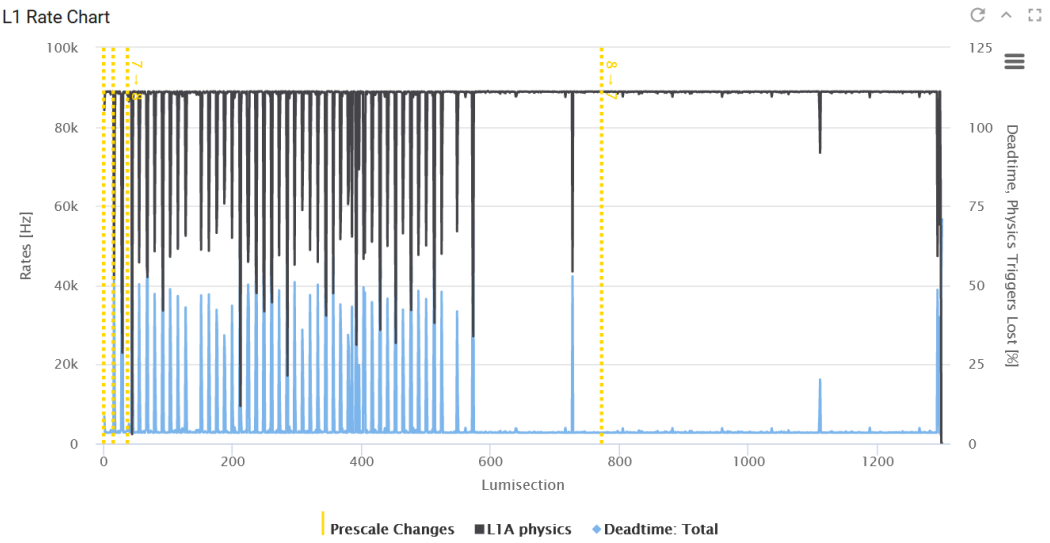


Figura 4.10: Ejemplo de *deadtimes* en el *trigger* L1.

La gráfica del error cuadrático medio también nos permite identificar aumentos anormales del error en intervalos de *lumisections* que se salen de la tendencia del ECM, como ocurre alrededor de la número 5500. Los datos de esta región han sido identificados como malos por los expertos, como se ve en la figura 4.8. Por tanto, vamos a tener dos tipos distintos de situaciones en las que vamos a etiquetar datos como malos usando el criterio del error cuadrático medio.

- ***Lumisections* cuyo ECM de reconstrucción constituya un máximo que se encuentre significativamente por encima del resto de datos**. Son los picos individuales o formados por unas pocas *lumisections* que se aprecian en las gráficas anteriores.

- **Conjuntos de *lumisections* que presenten un ECM superior a lo esperado según al aumento progresivo que se observa debido a la caída en la luminosidad**. Estas pueden presentar un error inferior a otras que se consideran como buenas, pero dicho error es significativamente superior a los que le rodean.

Para el primer tipo de datos malos **se establecerá un parámetro de corte para el modelo**, que no es más que un valor que permite descartar aquellas *lumisections* que tengan un error de reconstrucción superior a dicho parámetro. El criterio para determinarlo depende del grado de exigencia que se quiera imponer para etiquetar los datos como buenos. En este trabajo se va a tomar el valor correspondiente al mayor ECM que sea consistente con la disminución progresiva de la luminosidad instantánea, es decir, la cota superior para los datos que se encuentran agrupados en la franja principal.

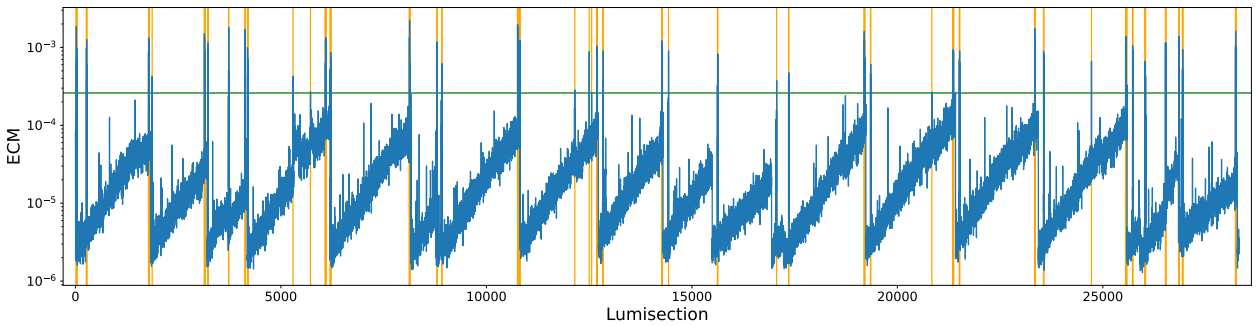


Figura 4.11: ECM de p_t en la era B con los valores aislados identificados como datos malos.

En la figura 4.11 se ha representado el parámetro de corte en verde, situado justo por encima de la zona donde se acumulan los datos cuyo ECM es razonable a la vista de la disminución progresiva de la luminosidad a lo largo del *fill*. En color naranja se han resaltado aquellos datos que el modelo etiquetará como malos por haberlo superado.

Para el segundo caso será necesario incluir un criterio diferente, ya que nos interesa medir la desviación respecto a la tendencia general del ECM. Para ello se va a utilizar una técnica conocida como **suavizado de series temporales**, un recurso que nos permite obtener una gráfica de los datos eliminando el ruido propio del comportamiento aleatorio.

En este trabajo se ha decidido comenzar empleando la media móvil, consistente en asociar a cada punto la media de un cierto número de datos que le rodean, conocidos como ventana.

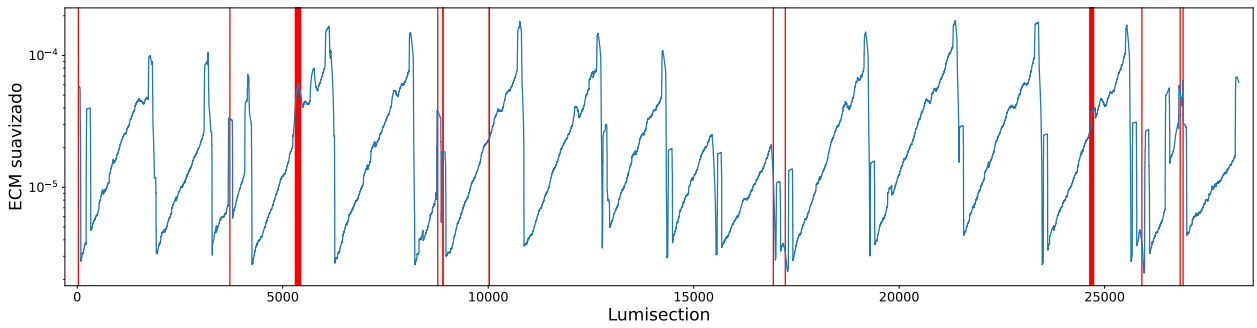


Figura 4.12: ECM de p_t en B suavizado con media móvil tomando una ventana de 100 datos.

La representación 4.12 nos permite comprobar que, efectivamente, las regiones que tienen un mayor número de datos etiquetados como malos por los expertos se corresponden con aumentos repentinos en el ECM suavizado respecto a la tendencia general. Estos cambios bruscos no pueden ser identificados utilizando un parámetro de corte al uso.

Sin embargo, hay que tener en cuenta que estamos trabajando con una distribución con numerosos valores atípicos, que se corresponden con las "agujas" que ya hemos tratado de etiquetar previamente como malas. Estos valores pueden provocar aumentos importantes en el ECM suavizado con la media a pesar de que el resto de datos de la ventana sean buenos. Así, como nuestro objetivo es únicamente detectar regiones en las que haya un número importante de datos que se salgan de la tendencia, podemos recurrir a una medida de tendencia central más robusta: la mediana. Esto quiere decir que su valor se ve menos afectado por la existencia de valores extremos, como se representa en la figura 4.13.

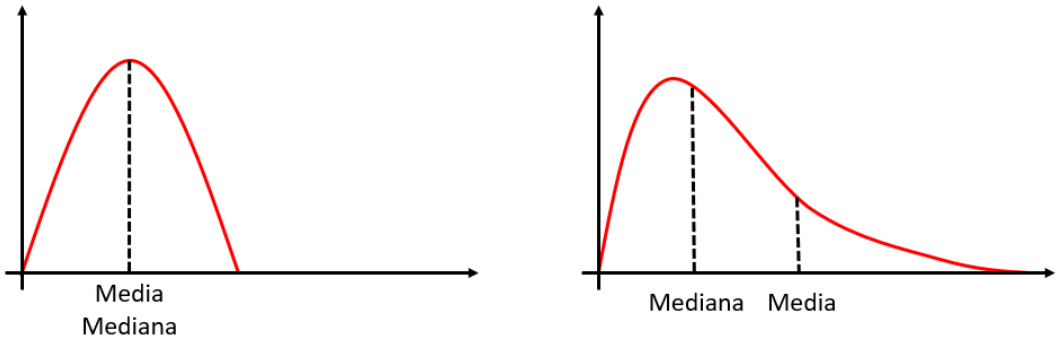


Figura 4.13: Ejemplo de robustez de la media y mediana. Fuente: Elaboración propia

La figura 4.14 muestra el ECM una vez que se le ha aplicado la mediana móvil. Se observa una función más suavizada, sin picos aislados producidos por valores atípicos y que aún así es capaz de identificar las regiones de *lumisections* que estamos buscando, como es el caso de la situada alrededor de la número 5500.

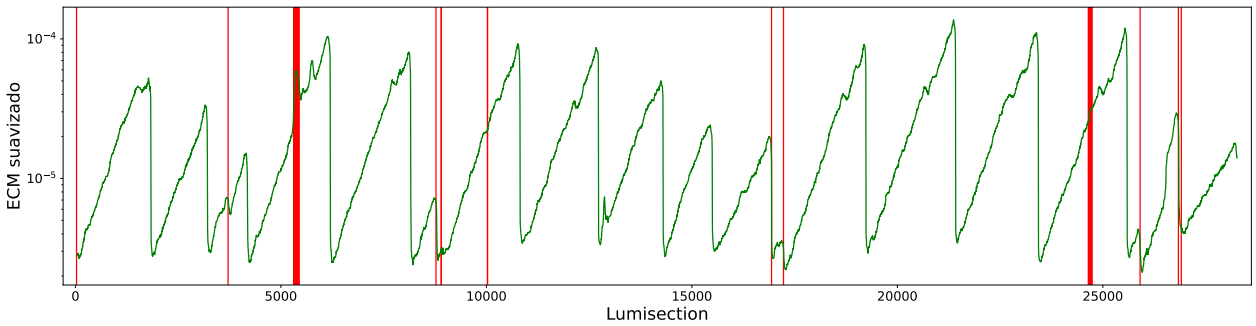


Figura 4.14: ECM de p_t en B suavizado con mediana móvil tomando una ventana de 100 datos.

La dificultad se encuentra en este punto en determinar cómo identificar estas regiones. Para ello la primera tarea ha sido determinar los máximos en la función suavizada y descartar aquellos que coinciden con los máximos debidos al final de los *fills*. A partir de ahí se incluye un criterio adicional consistente en imponer cuántas *lumisections* debe tener como mínimo el pico para considerarlo, tomando en este trabajo un valor de 30. En la programación del modelo también se ha impuesto un criterio sobre la prominencia mínima que debe tener el pico. Las zonas descartadas por este método se recogen en color naranja en la figura 4.15. Es importante tener en cuenta que al estar utilizando una escala logarítmica, los picos de la parte inferior de la gráfica aparentan ser más altos que los de la parte superior en igualdad de condiciones.

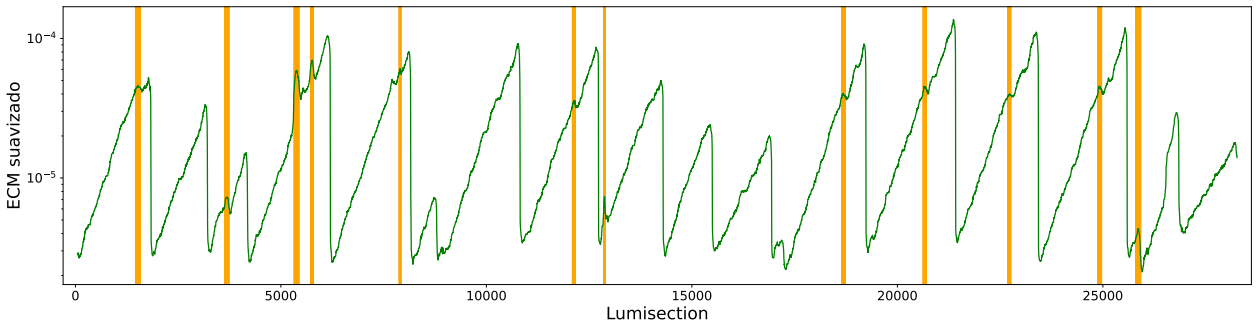


Figura 4.15: ECM de p_t en B suavizado con las regiones identificadas como datos malos.

El proceso final pasa por **juntar los datos descartados por ambos motivos**, algo que permite identificar todos aquellos que han sido etiquetados como malos por el modelo para la propiedad física estudiada, que en este ejemplo es el momento transverso.

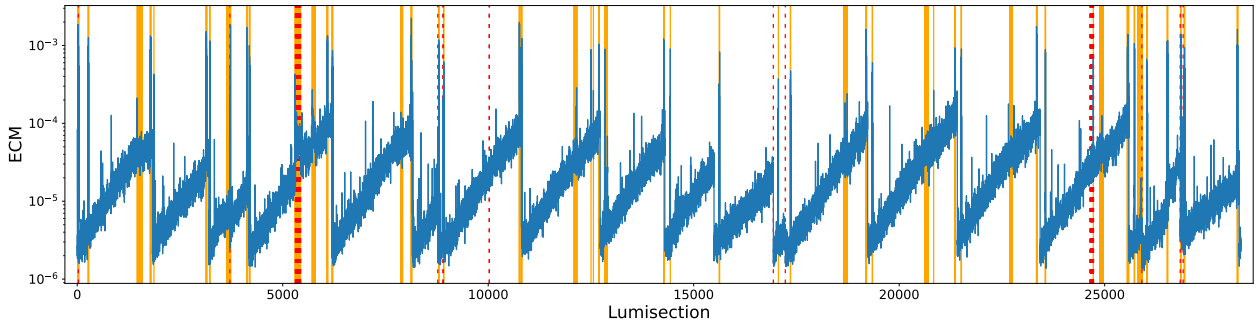


Figura 4.16: Datos etiquetados como malos por el modelo (naranja) y los expertos (rojo).

Como disponemos de las distribuciones de cuatro magnitudes físicas de los muones registrados, este procedimiento se repetiría para todas ellas, considerando que **una *lumisection* es mala si es etiquetada como tal en alguna de ellas** (equivalentemente, solo será buena si lo es para las cuatro magnitudes). Esto permite obtener una única etiqueta para cada *lumisection*. Es decir, la etiqueta de *lumisection* mala se asigna mediante la operación lógica OR.

La implementación técnica de este modelo basado en PCA y el estudio de su comportamiento ha sido realizado en primera persona y de forma original haciendo uso del lenguaje de programación *Python* y los datos de muones proporcionados en formato CSV. Los detalles relacionados con el código se pueden encontrar en el anexo B.

Capítulo 5

Resultados

5.1. Número de componentes principales

Una de las primeras cuestiones que nos podemos plantear a la hora de trabajar con PCA es el número de componentes principales con las que nos tenemos que quedar. Con el objetivo de llevar a cabo un estudio de este hecho se ha realizado el entrenamiento en la era C y la evaluación en la era B considerando un número de componentes entre 1 y 100 para cada propiedad y **se han calculado los coeficientes $\mathbf{F_{1/2}}$ y GM, que se tratarán de maximizar**. Se ha considerado la evaluación en la era B porque se ha podido comprobar en el capítulo 4 que presenta ambos tipos de *lumisections* de ECM elevado, tanto aisladas como agrupadas.

La razón por la que se recurre a utilizar $F_{1/2}$ en lugar de F_1 es el interés en dar mayor peso al PPV frente al TPR. Esto es debido a que el primero nos permite controlar el número de falsos positivos, mientras que el segundo controla el de falsos negativos, siendo en nuestro caso preferible descartar alguna *lumisection* buena que dejar pasar *lumisections* malas, ya que pueden afectar a la calidad de las conclusiones físicas que se extraigan en las siguientes fases del proceso. Por su parte, la razón por la que se utiliza la media geométrica GM es la necesidad de tener una medida del comportamiento del modelo que tenga en cuenta el número de verdaderos negativos.

En este estudio nos centramos en un análisis principalmente cualitativo, fijándonos principalmente en la tendencia general de las curvas de la figura 5.1 y no tanto en las fluctuaciones concretas que puedan aparecer, pues estas se deben a las particularidades de los conjuntos de entrenamiento y evaluación empleados.

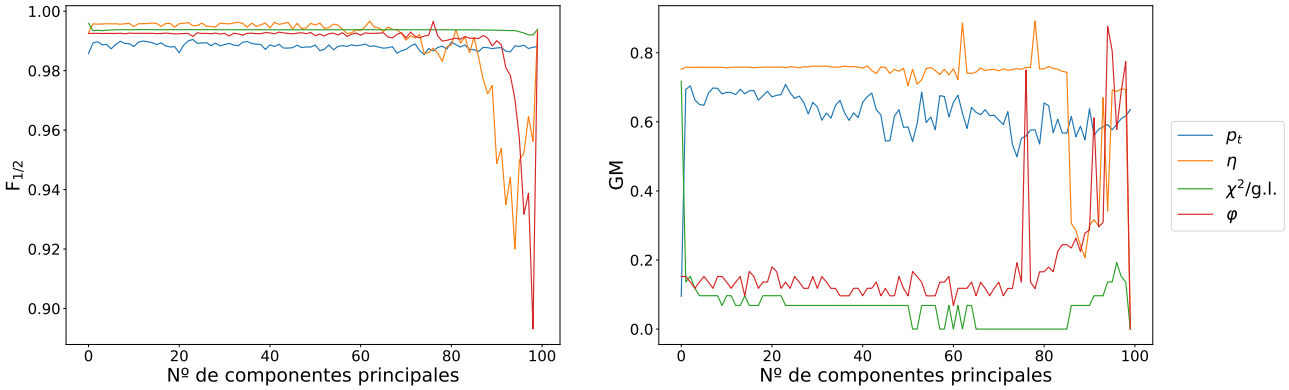


Figura 5.1: Estudio del número de componentes necesarias para cada propiedad.

Comenzamos estudiando las propiedades η, χ^2/g.l. y φ, que presentan gráficas análogas. En todas ellas se observa un comportamiento ruidoso para un número alto de componentes principales, produciéndose una caída del valor de $F_{1/2}$ y un aumento de GM, aunque muchas veces con un gran número de fluctuaciones. La razón para ello se encuentra en que un número alto de PC lleva a una mejor reconstrucción de todos los histogramas, de manera que resulta más difícil emplear las diferencias entre los originales y los reconstruidos como criterio de clasificación.

Por debajo de unas 70 componentes principales, el comportamiento es aproximadamente constante, exceptuando la reconstrucción con una única componente principal, que tiene un rendimiento ligeramente inferior en ambos coeficientes $F_{1/2}$ y GM para η y φ y superior en el caso de χ^2/g.l. .

Si consideramos el momento transverso p_t, el coeficiente $\mathrm{F}_{1/2}$ es aproximadamente constante, mientras que en el caso de GM hay una tendencia general decreciente. Usando una reconstrucción con una o dos componentes principales se obtiene un rendimiento inferior. Esto puede deberse a una mayor variabilidad entre histogramas debida al prescalado del *trigger*.

Con todas las consideraciones presentadas previamente para cada una de las propiedades físicas estudiadas, se ha decidido tomar **el siguiente número de componentes principales**:

Propiedad	p_t	φ	χ^2/g.l.	η
Componentes	3	2	1	2

Cuadro 5.1: Número de componentes principales para cada propiedad.

5.2. Evaluación del modelo

Se realiza el **entrenamiento con los datos de tres de las eras** y la **evaluación en la era restante**, recogiendo la matriz de confusión y los parámetros de resumen. Además, se incluye una gráfica que ilustre algún comportamiento interesante del modelo en dicha era usando el ECM y las etiquetas de alguna de las propiedades en un intervalo reducido de *lumisections*. Las gráficas completas para cada propiedad física y cada era se incluyen en el anexo A.

A continuación se mostrará el valor del ECM para cada *bin* en lugar de cada histograma usando **todas las *lumisections* etiquetadas como malas por el modelo** en esa era. En caso de funcionamiento óptimo del detector, la gráfica debería tener el aspecto ideal mostrado en la figura 4.5. Una desviación de ella nos daría una idea de que existe una región del detector que está fallando. Se ha usado η por dar una idea geométrica de CMS y ser la propiedad más relevante en la detección de anomalías, como se verá en la siguiente sección. La última gráfica muestra un ejemplo de reconstrucción del histograma de η para una *lumisection* etiquetada como mala y que ilustra algún posible fallo que se haya apreciado en la figura anterior.

Evaluación en la era A. Entrenamiento en B, C y D.

		Predicción modelo	
		Bueno	**Malo**
Valor experto	**Bueno**	49184	1831
	Malo	171	76

$$\text{PPV} = 0.9965$$
$$\text{TPR} = 0.9641$$
$$\text{TNR} = 0.3077$$

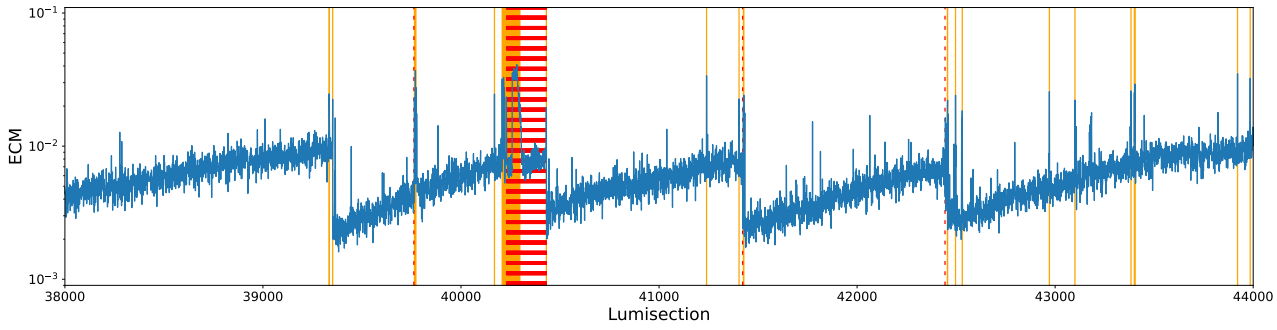


Figura 5.2: *Lumisections* buenas en un *run* etiquetado como malo identificadas con η.

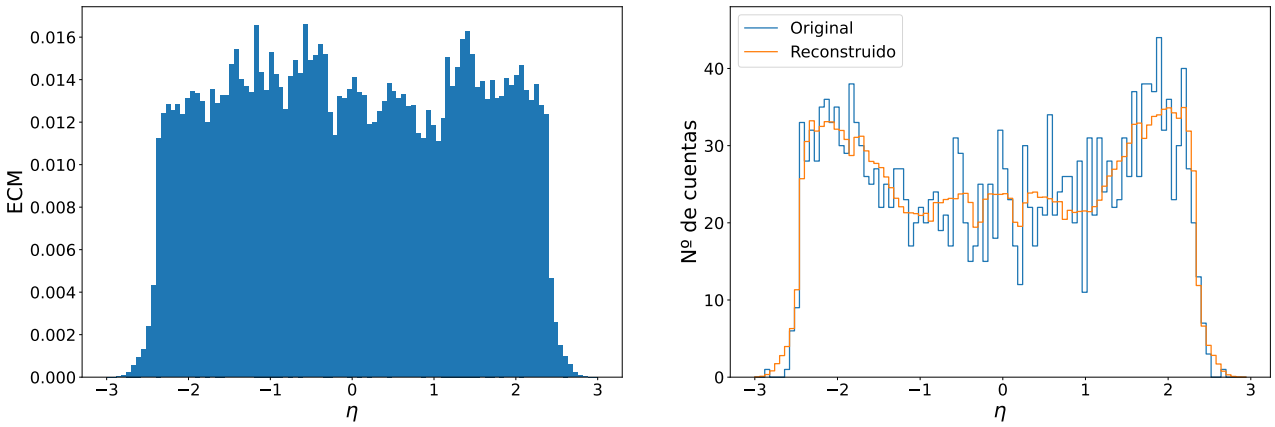


Figura 5.3: ECM en cada *bin* en la era A y ejemplo de *lumisection* etiquetada como mala.

En la evaluación en esta era se obtienen muy buenos resultados en la clasificación de los datos buenos como tal, algo que se puede observar directamente en la matriz de confusión. Su identificación de datos malos ha sido más discreta, obteniendo un TNR igual a 0.3077. Una parte de esto se debe a la identificación de *lumisections* buenas en *runs* etiquetados como malos por los expertos, como se ve en la figura 5.2. Recordamos que las *lumisections* en naranja han sido etiquetadas como malas por el modelo y las rojas, por los expertos.

La gráfica izquierda de la figura 5.3 muestra una menor error de reconstrucción para los dos picos situados en torno a $\eta = \pm 2$, lo que indica que los principales problemas en la detección se producen a pseudorapidez baja. Un ejemplo de ello se ve en las diferencias entre el histograma original y reconstruido de la gráfica de la derecha.

Evaluación en la era B. Entrenamiento en A, C y D.

		Predicción modelo	
		Bueno	**Malo**
Valor experto	**Bueno**	27024	1076
	Malo	87	127

PPV = 0.9968
TPR = 0.9617
TNR = 0.5935

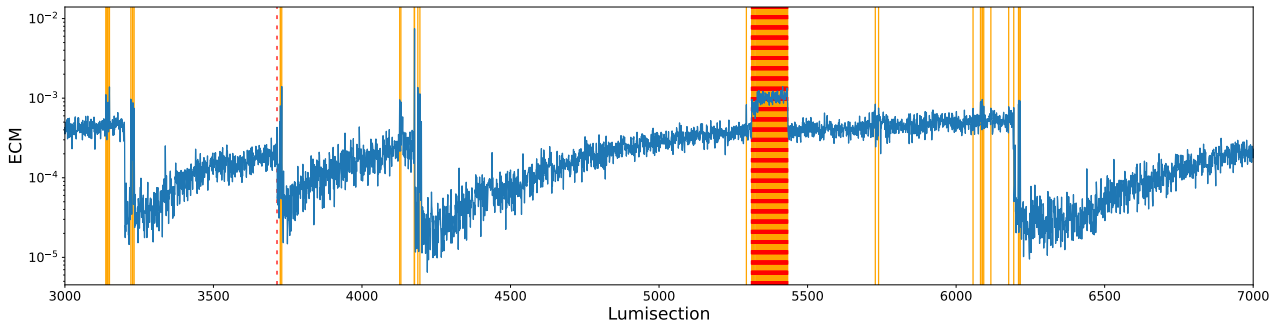


Figura 5.4: Identificación de una región anómala usando el ECM de χ^2/g.l. en la era B.

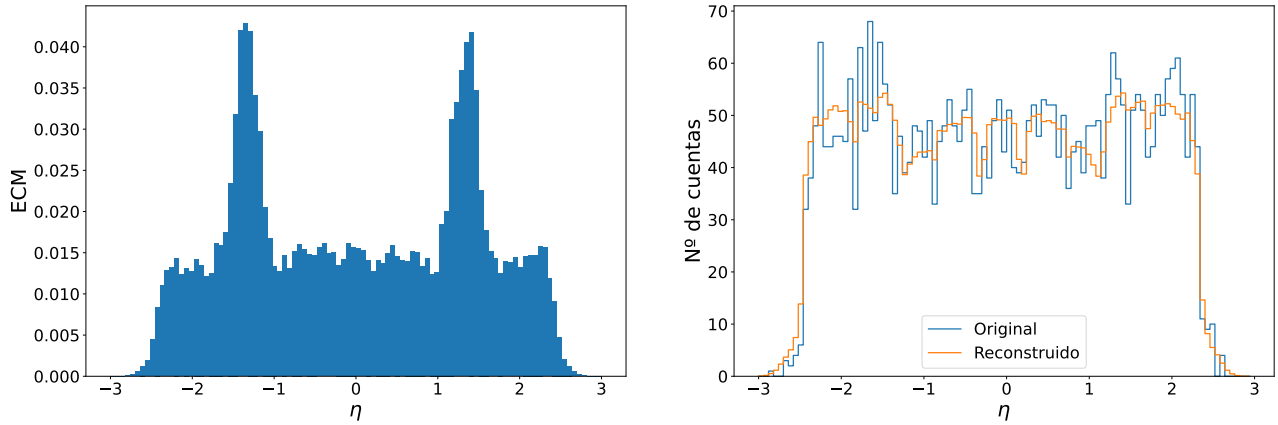


Figura 5.5: ECM en cada *bin* en la era B y ejemplo de *lumisection* etiquetada como mala..

El comportamiento del modelo en la era B ha sido más satisfactorio, alcanzando unas medidas resumen más cercanas a 1. La principal fuente de falsos positivos ha sido la existencia de *runs* etiquetados completamente como malos en los que el modelo ha seleccionado una parte de las *lumisections* como buenas.

La gráfica del ECM en la figura 5.5 muestra que existen dos regiones, en torno a $\eta = \pm 1.5$ que presentan un ECM en la reconstrucción mucho más alto que las demás en los datos que el modelo ha clasificado como malos. Esto podría indicar que existe algún error en los subdetectores situados en esas direcciones. El ejemplo de su derecha ilustra que las detecciones en esta era para η bajo se ajustaron mucho más a lo esperado que aquellas correspondientes a los dos máximos de los extremos.

Evaluación en la era C. Entrenamiento en A, B y D

		Predicción modelo	
		Bueno	**Malo**
Valor experto	**Bueno**	24952	1329
	Malo	62	333

$$\text{PPV} = 0.9975$$
$$\text{TPR} = 0.9494$$
$$\text{TNR} = 0.8430$$

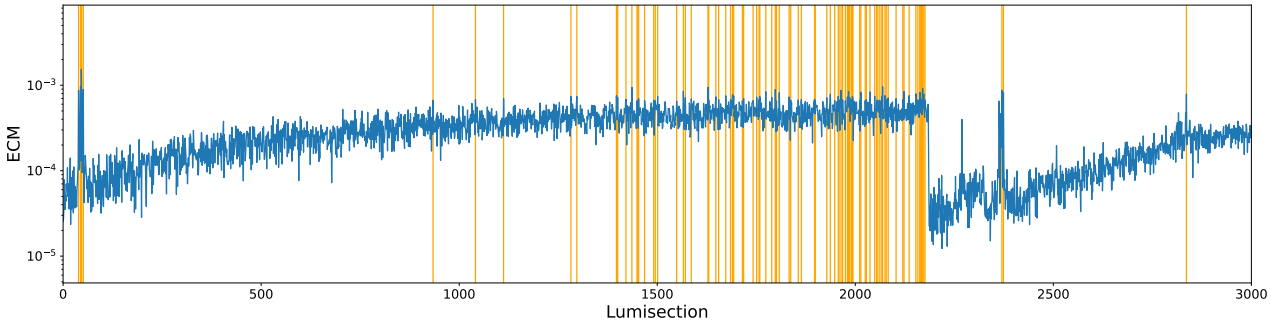


Figura 5.6: Acumulación de datos etiquetados como malos por el ECM de χ^2/g.l. en la era C.

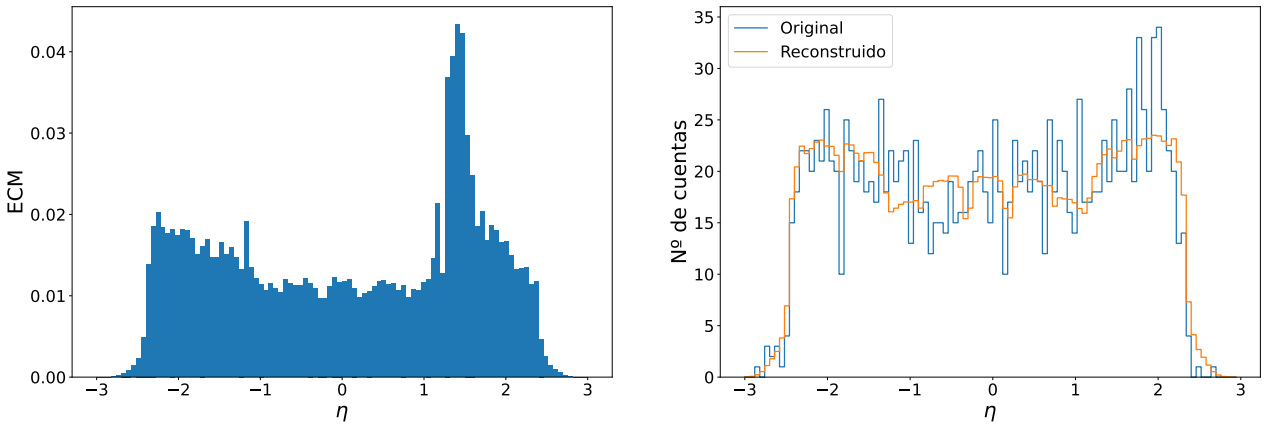


Figura 5.7: ECM en cada *bin* en la era C y ejemplo de *lumisection* etiquetada como mala.

Los resultados en la era C son realmente buenos, alcanzando valores significativamente altos de PPV, TPR y TNR. Buena parte de los falsos negativos provienen de las primeras 3000 *lumisections*, donde el modelo ha señalado varias regiones con posibles errores debido al ECM de χ^2/g.l., como se puede ver en la figura 5.6.

La representación del ECM en la figura 5.7 vuelve a ser relevante para la identificación de la región del detector que puede estar presentando problemas, siendo $\eta = 1.5$ en el caso de esta era. El ejemplo de su derecha ilustra una *lumisection* con dicha problemática, en la que el contaje de muones en uno de los tapones del detector es mucho mayor que en el otro.

Evaluación en la era D. Entrenamiento en A, B y C.

		Predicción modelo	
		Bueno	**Malo**
Valor experto	**Bueno**	119007	4049
	Malo	443	133

PPV = 0.9963
TPR = 0.9671
TNR = 0.2309

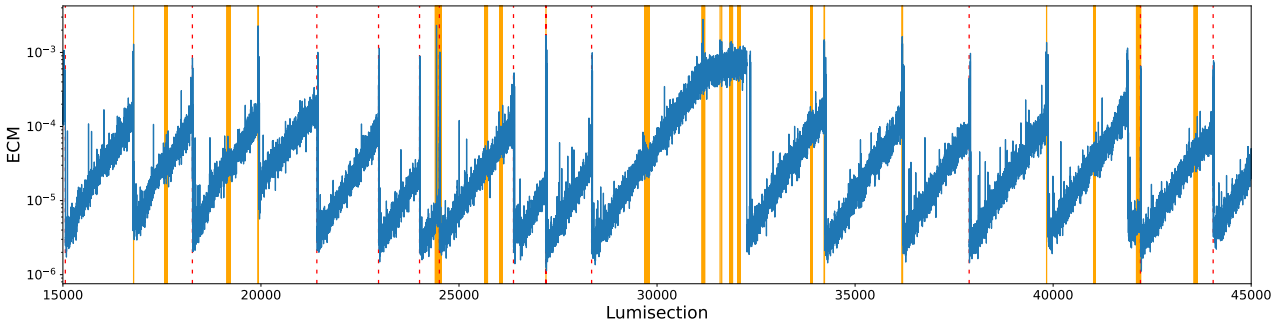


Figura 5.8: Región con un error superior a lo esperado en el ECM de p_t en la era D.

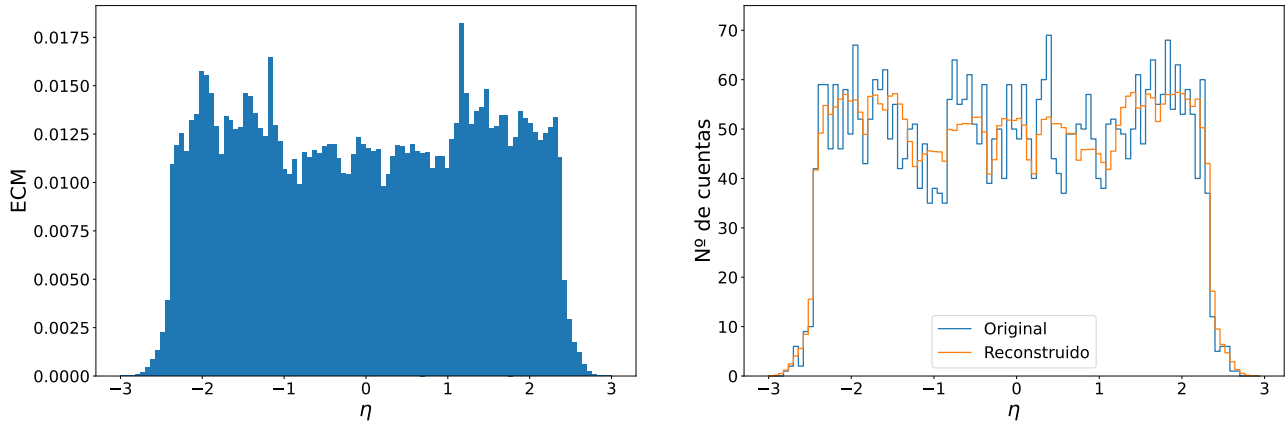


Figura 5.9: ECM en cada *bin* en la era D y ejemplo de *lumisection* etiquetada como mala.

El comportamiento en la era D vuelve a ser algo peor al comparar con las etiquetas de los expertos, ya que el modelo ha sido bastante conservador en la clasificación de *lumisections* como malas. La explicación pasa por notar que el alto ECM alcanzado en torno a la *lumisection* 32000 y que se recoge en la figura 5.8 ha provocado que el parámetro de corte tenga que ser relativamente alto para evitar un aumento excesivo de los falsos negativos en las cuatro propiedades.

Las gráficas de la figura 5.9 muestran un comportamiento con pocas desviaciones del caso ideal, mostrando únicamente una mejor reconstrucción para muones con alta pseudorapidez.

Considerando todas las eras, se observa que **el modelo ha presentado un comportamiento positivo**, si bien el número de falsos positivos y negativos es relativamente alto en algunos casos. La explicación para este hecho se encuentra en la forma en la que se han asignado las etiquetas de los expertos, ya que este proceso se realiza *run* a *run* en lugar de considerar *lumisections* individuales, como hace el modelo desarrollado en este trabajo, lo que hace que su uso haya permitido identificar *lumisections* buenas en *runs* etiquetados como malos y viceversa.

Por otro lado, es importante considerar que en el conjunto de datos empleado únicamente estamos trabajando con cuatro propiedades físicas relativas a los muones detectados en CMS, lo que supone una cantidad de información muy restringida respecto a la manejada por el DQM en la tarea de certificación, para la cual cuentan con datos referentes a las propiedades de otras partículas, el comportamiento del detector, del *trigger* o de los sistemas informáticos, entre otros.

A pesar de todo ello, se puede ver que el comportamiento del modelo a la hora de identificar posibles *lumisections* con datos malos es satisfactorio, especialmente si tenemos en cuenta que estamos trabajando con un método consistente en resolver un problema de autovalores. En cualquier caso, las ideas de aplicación de aprendizaje automático no pretenden por el momento sustituir el trabajo de las personas encargadas de la certificación, sino reducir su carga de trabajo y permitir una mayor granularidad, sirviendo como un complemento a los métodos empleados.

5.3. Relevancia de las propiedades físicas

La clasificación de datos como buenos o malos con este modelo se ha hecho en función de la realizada para cada propiedad de los muones con la operación OR. Es interesante estudiar cuál ha sido el comportamiento de cada una, determinando cuáles son las que permiten discriminar mejor y cuáles no aportan información relevante a efectos de la certificación con este modelo.

Para cada propiedad se ha supuesto que ella **es el único criterio para etiquetar las *lumisections*** como buenas o malas y se ha determinado la matriz de confusión. Esto se ha repetido para la evaluación en cada una de las eras (con el entrenamiento en las tres restantes) y se han sumado las matrices para obtener el comportamiento general a lo largo de las cuatro eras.

Momento transverso $\mathbf{p_t}$

		Predicción modelo	
		Bueno	**Malo**
Valor experto	**Bueno**	222044	6408
	Malo	1223	209

Pseudorapidez $\boldsymbol{\eta}$

		Predicción modelo	
		Bueno	**Malo**
Valor experto	**Bueno**	227679	773
	Malo	834	598

Ángulo azimutal $\boldsymbol{\varphi}$

		Predicción modelo	
		Bueno	**Malo**
Valor experto	**Bueno**	227958	494
	Malo	1422	10

Coeficiente de ajuste $\boldsymbol{\chi^2}$/g.l.

		Predicción modelo	
		Bueno	**Malo**
Valor experto	**Bueno**	226452	2000
	Malo	1316	116

Cuadro 5.2: Matrices de confusión evaluando en las cuatro eras para cada propiedad.

De las matrices anteriores se pueden extraer algunas conclusiones. El uso de p_t lleva a la identificación de muchas *lumisections* como malas, principalmente debidas al criterio de descartar regiones cuyo ECM se desvíe de lo esperado según la caída progresiva de la luminosidad. **La pseudorapidez es la propiedad que mejor se ajusta a las etiquetas de los expertos**, mostrando la importancia de tener en cuenta la geometría del detector y siendo una propiedad ampliamente utilizada por los trabajadores del DQM en la certificación de datos.

Las magnitudes φ y χ^2/g.l. son las que obtienen resultados más discretos si se comparan con los valores de los expertos. De hecho, en el caso del ángulo azimutal se puede comprobar en la segunda columna de la matriz que el modelo apenas marca *lumisections* como malas.

Se podría llegar a plantear en algunas situaciones el uso de η como única propiedad a tener en cuenta por el modelo, pues permite obtener un TPR = 0.9966 y un TNR = 0.4176, frente al TPR = 0.9637 y TNR = 0.4672 que se obtiene al usar las cuatro magnitudes. Esto se interpreta como que el modelo que únicamente usa la pseudorapidez detecta un número ligeramente inferior de datos etiquetados como malos pero lo hace de forma muy selectiva, reduciendo considerablemente el número de falsos negativos.

5.4. Mediana y media en la búsqueda de regiones anómalas

Durante la presentación del modelo de aprendizaje automático en el capítulo 4 se propuso el uso de la mediana móvil frente a la media móvil para la identificación de los grupos de *lumisections* con un ECM superior al que les correspondía según la luminosidad. La justificación para dicha elección se basó en que era una medida más robusta, de manera que no se vería afectada por las *lumisections* individuales con un error de reconstrucción muy elevado.

Con el fin de comprobar que esta elección se traduce en un mejor comportamiento del modelo completo, se ha seguido un procedimiento análogo al del apartado anterior, en el que se realiza la evaluación en cada era después de hacer el entrenamiento con las tres eras restantes y luego se suman las cuatro matrices de confusión. Esto se traduce en los siguientes resultados:

Mediana móvil

		Predicción modelo	
		Bueno	**Malo**
Valor experto	**Bueno**	220167	8285
	Malo	763	669

Media móvil

		Predicción modelo	
		Bueno	**Malo**
Valor experto	**Bueno**	207983	20469
	Malo	1005	427

En las matrices de confusión se observa claramente un **mejor comportamiento al hacer uso de la mediana móvil**, ya que su desempeño es superior tanto para la identificación de *lumisections* buenas como malas. Esto se puede comprobar también a través de la fracción de verdaderos positivos, obteniendo TPR = 0.9637 para la mediana y TPR = 0.9104 para la media, o la fracción de verdaderos negativos, siendo esta TNR = 0.467 para la mediana y TNR = 0.298 para la media.

Conclusiones

El enorme volumen de datos generados en el detector CMS del LHC supone un reto para la certificación de datos debido a que los métodos empleados actualmente suponen la revisión manual de cientos de histogramas, con el consecuente aumento de la probabilidad de errores humanos y la necesidad de realizar dicha clasificación de los datos en buenos o malos por *runs*. En este contexto surge el interés por aplicar técnicas de aprendizaje automático que permitan aliviar la carga de trabajo y aumentar la fiabilidad y tiempo de respuesta de la certificación.

En este trabajo se ha presentado un modelo de aprendizaje semisupervisado basado en el análisis de componentes principales. Se ha comprobado que esta técnica, habitualmente empleada para reducir la dimensionalidad, tiene un gran potencial en la detección de anomalías a través del error de reconstrucción. Este modelo se ha diseñado, entrenado y evaluado teniendo en cuenta los diversos inconvenientes del problema planteado: el desequilibrio entre datos buenos y malos, las grandes variaciones estadísticas, el enorme volumen de datos y la aproximación de asociar las etiquetas de los *runs* de los expertos a las *lumisections* que los forman.

La utilización de un doble criterio de descarte, uno de ellos basado en un parámetro de corte y el otro en las desviaciones respecto a la tendencia general, ha permitido buscar un equilibrio entre ambos que garantice que ninguno de los dos tenga que ser muy exigente, lo que permite un descenso del número de falsos negativos respecto al uso individual de uno de ellos. Se ha comprobado también la eficiencia de la mediana móvil en la identificación de regiones anómalas.

El modelo diseñado ha permitido obtener una mayor granularidad, pudiendo realizar la certificación de forma individual para cada *lumisection* en lugar de cada *run*. Un estudio del error cuadrático medio ha permitido también determinar cuáles son las regiones espaciales del detector causantes de la aparición de datos malos en cada era. Además de ello, se ha comprobado que la pseudorapidez es habitualmente la propiedad de los muones cuyo estudio proporciona más información en términos de certificación.

Bibliografía

[1] R. Mann. *An introduction to particle physics and the standard model*. 1.ª ed. Taylor & Francis, 2010.

[2] B. Povh. *Particles and Nuclei*. 7.ª ed. Springer, 2015.

[3] R.L. Workman et al (Particle Data Group). *Review of Particle Physics*. Vol. 2022. PTEP, 2022.

[4] *Standard Model of Elementary Particles*. `https://commons.wikimedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg`. Visitada el 15-02-2023.

[5] E. Daw. *Lecture 7 - Rapidity and Pseudorapidity*. University of Sheffield. 2012.

[6] *CMS coordinate system*. `https://tikz.net/axis3d_cms/`. Visitada el 15-02-2023.

[7] *CERN*. `https://home.cern/`. Visitada el 28-03-2023.

[8] Education, Communications and Outreach Group. *LHC the guide FAQ*. CERN, 2021.

[9] *Large Hadron Collider Timeline*. `https://timeline.web.cern.ch/timeline-header/93`. Visitada el 26-03-2023.

[10] O. Aberle et al. *High-Luminosity Large Hadron Collider (HL-LHC): Technical design report*. CERN Yellow Reports: Monographs. CERN, 2020.

[11] *Longer term LHC schedule*. `http://lhc-commissioning.web.cern.ch/schedule/LHC-long-term.htm`. Visitada el 28-03-2023.

[12] *CMS*. `https://cms.cern/`. Visitada el 28-03-2023.

[13] CMS Collaboration. *CMS Physics : Technical Design Report Volume 1: Detector Performance and Software*. Inf. téc. CERN, 2006.

[14] The CMS Collaboration. *The CMS experiment at the CERN LHC*. Journal of Instrumentation, 2008.

[15] D. South y M. Turcato. *Review of Searches for Rare Processes and Physics Beyond the Standard Model at HERA*. Vol. 76. The European Physical Journal C, 2016.

[16] M. Ressegotti. *Overview of the CMS Detector Performance at LHC Run 2*. Vol. 5. Universe, 2019.

[17] *CMS Tracker Detector Performance Results*. `https://twiki.cern.ch/twiki/bin/view/CMSPublic/DPGResultsTRK`. Visitada el 02-06-2023.

[18] CMS Collaboration. *The CMS trigger system*. Vol. 12. Journal of Instrumentation, 2017.

[19] J. C. Béjar. *Computación Grid para el experimento CMS del LHC*. Tesis doctoral, Universidad Complutense de Madrid, 2007.

[20] V. Azzolini et al. *The Data Quality Monitoring Software for the CMS experiment at the LHC: past, present and future*. EPJ Web of Conferences, 2019.

[21] L. Tuura et al. *CMS data quality monitoring: Systems and experiences*. Journal of Physics: Conference Series, 2010.

[22] O. Gutsche. *Validation of Software Releases for CMS*. Vol. 219. Journal of Physics: Conference Series, 2010.

[23] *Historic DQM for the CMS/LHC Tracker*. `https://indico.cern.ch/event/680759/contributions/2789265/attachments/1558061/2451216/HDQM_at_INPP.pdf`. Visitada el 02-06-2023.

[24] E. Alpaydin. *Introduction to machine learning*. 3.ª ed. MIT press, 2020.

[25] S. J. Russel y P. Norvig. *Artificial Intelligence: A Modern Approach*. 3.ª ed. Prentince Hall, 2010.

[26] M. R. Carbone. *When not to use machine learning: A perspective on potential and limitations*. MRS Bulletin, 2022.

[27] *Overfitting in Machine Learning*. `https://h2o.ai/wiki/overfitting/`. Visitada el 03-06-2023.

[28] I. Martín de Diego et al. *General Performance Score for classifcation problems*. Applied Intelligence, 2021.

[29] M. Borisyak et al. *Towards automation of data quality system for CERN CMS experiment*. Journal of Physics: Conference Series, 2017.

[30] Albertsson et al. *Machine learning in high energy physics community white paper*. Publisher of Physics: Conference Series, 2018.

[31] V. Wachirapusitan. *Machine Learning applications for Data Quality Monitoring and Data Certification within CMS*. Journal of Physics: Conference Series, 2023.

[32] M. Alcalde Martínez. *Estudio con aprendizaje automático de la certificación de datos físicos del detector CMS de LHC (CERN) utilizando Non-Negative-Matrix-Factorization (NMF)*. Trabajo Fin de Grado, Universidad de Oviedo, 2022.

[33] H. Calvo Castro. *Estudio con aprendizaje automático de la certificación de datos físicos del detector CMS de LHC (CERN) utilizando Auto Encoders (AEs)*. Trabajo Fin de Grado, Universidad de Oviedo, 2022.

[34] J. Shlens. *A tutorial on principal component analysis*. arXiv:1404.1100, 2014.

[35] I. T. Jolliffe. *Principal Component Analysis*. 2.ª ed. Springer, 2002.

[36] I. T. Jolliffe y J. Cadima. *Principal component analysis: a review and recent developments*. Phil. Trans. R. Soc. A, 2016.

[37] R. Deserio. *Statistical Analysis of Data in the Linear Regime*. University of Florida - Department of Physics, 2015.

[38] G. James et al. *An Introduction to Statistical Learning*. 1.ª ed. Springer, 2013.

Anexo A: Comportamiento del modelo en cada magnitud física

Este anexo recoge las gráficas en las que se representa el ECM para cada propiedad física de los muones y en cada una de las eras. Se recogen también las etiquetas asignadas por los expertos del DQM como líneas punteadas de color rojo y las asignadas por el modelo debido a la magnitud que se está representando en naranja. Es decir, las *lumisections* resaltadas en color naranja en cada gráfica son las que se han etiquetado como malas atendiendo únicamente a la magnitud física que se está representando, antes por tanto de aplicar la operación lógica OR.

El objetivo es que el lector pueda disponer de estas gráficas a modo de consulta en caso de ser necesario para comprender mejor los resultados recogidos en el capítulo 5 de este trabajo. Pueden resultar de interés para comprobar qué magnitudes físicas han sido capaces de identificar cada una de las anomalías de los datos y comprobar los resultados generalizados para las cuatro eras en la sección 5.3 para el caso particular de cada era.

A la hora de interpretar las gráficas es importante notar que las líneas rojas y naranjas son más anchas de lo que correspondería a una única *lumisection*, ya que si su anchura se representase a escala no serían visibles debido a la medida del eje de abscisas. La consecuencia de este fenómeno es que puede parecer a primera vista que el modelo está etiquetando más *lumisections* como malas de lo que realmente se refleja en las matrices de confusión.

Evaluación en la era A. Entrenamiento en B, C y D.

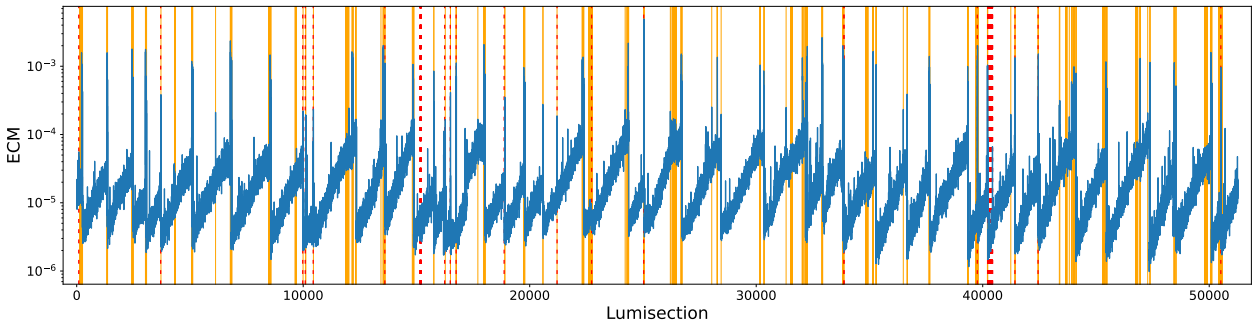


Figura A.1: ECM y etiquetas asignadas de p_t en la era A.

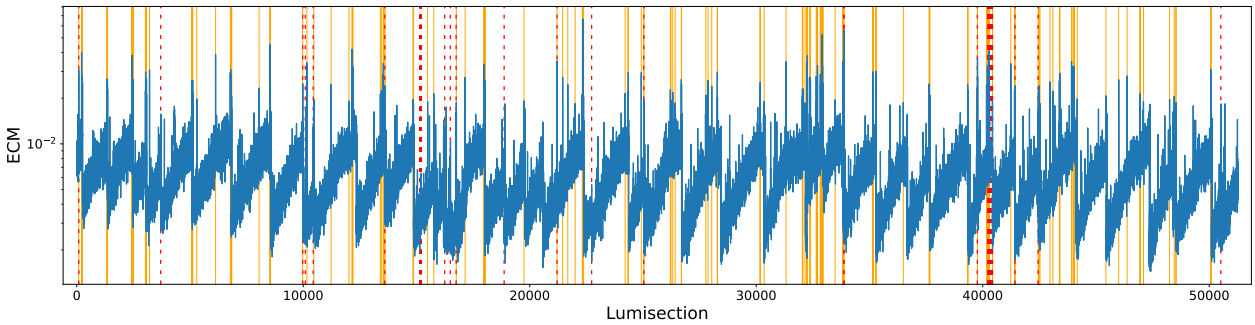


Figura A.2: ECM y etiquetas asignadas de η en la era A.

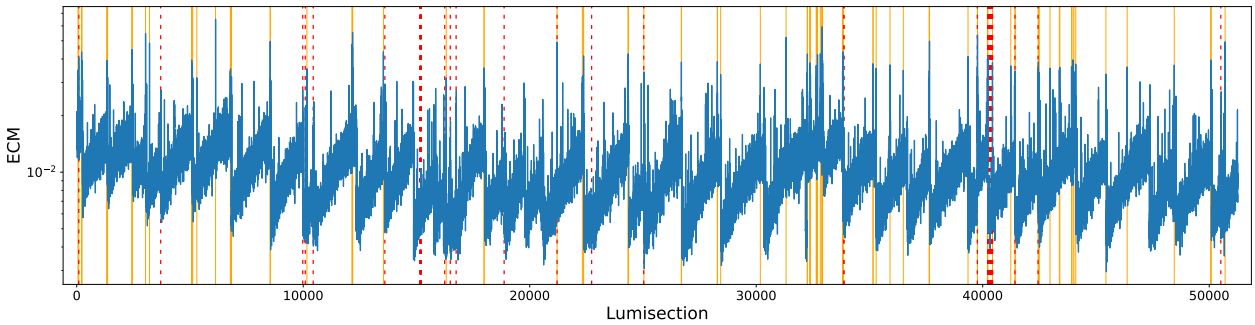


Figura A.3: ECM y etiquetas asignadas de φ en la era A.

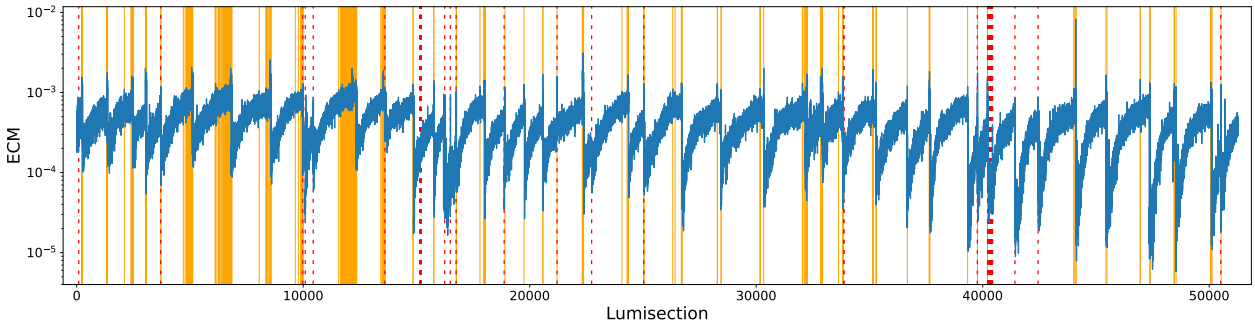


Figura A.4: ECM y etiquetas asignadas de χ^2/g.l. en la era A.

Evaluación en la era B. Entrenamiento en A, C y D.

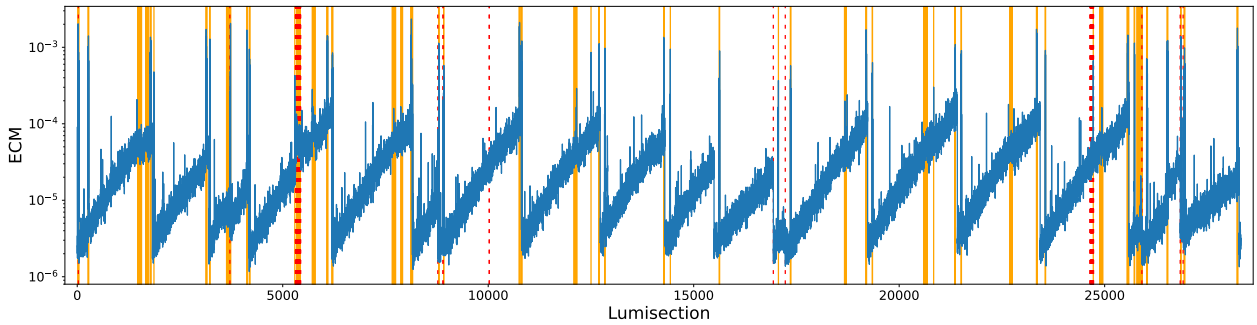


Figura A.5: ECM y etiquetas asignadas de p_t en la era B.

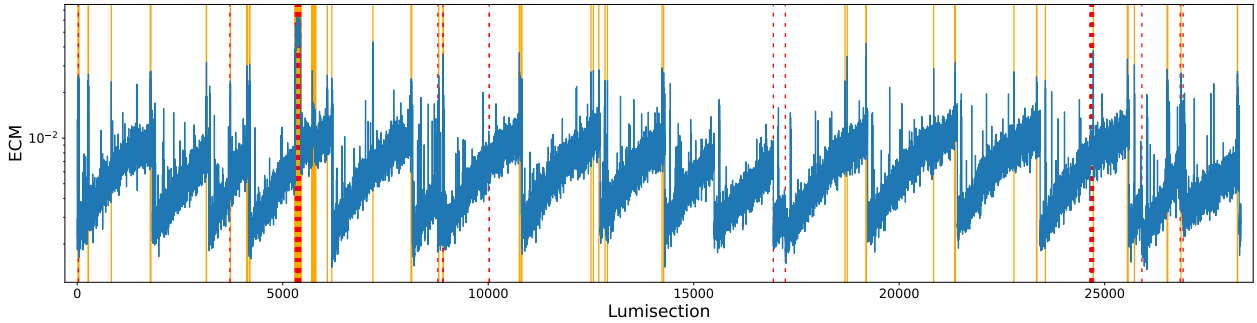


Figura A.6: ECM y etiquetas asignadas de η en la era B.

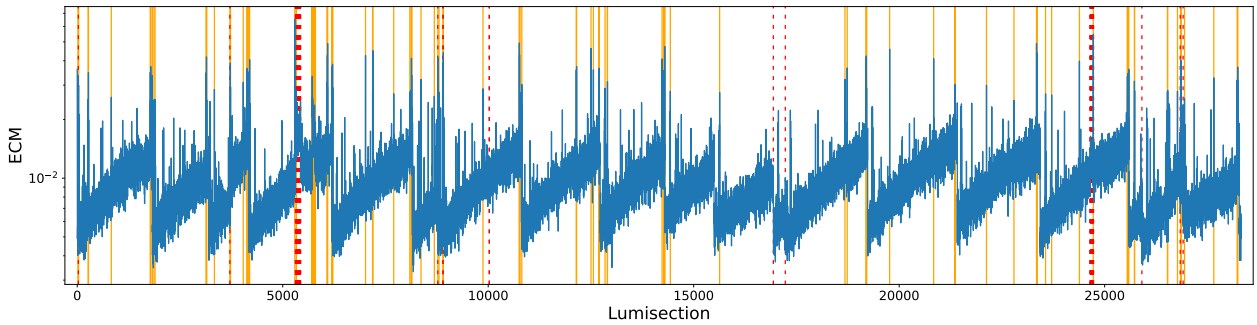


Figura A.7: ECM y etiquetas asignadas de φ en la era B.

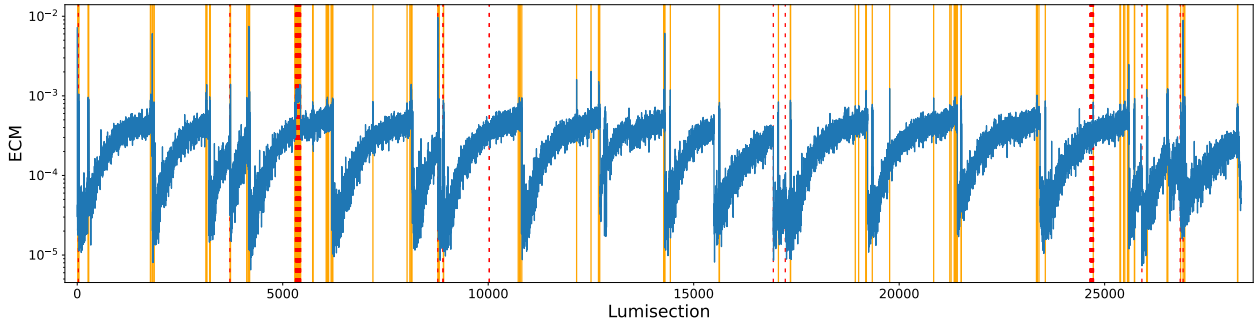


Figura A.8: ECM y etiquetas asignadas de χ^2/g.l. en la era B.

Evaluación en la era C. Entrenamiento en A, B y D.

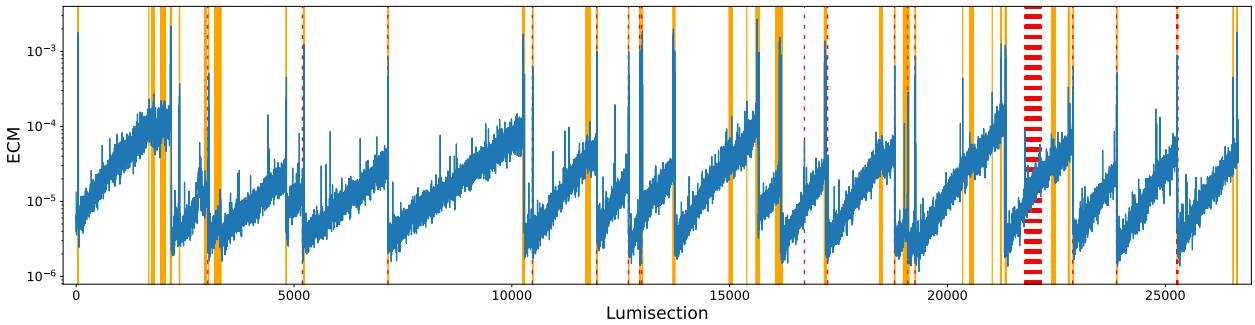


Figura A.9: ECM y etiquetas asignadas de p_t en la era C.

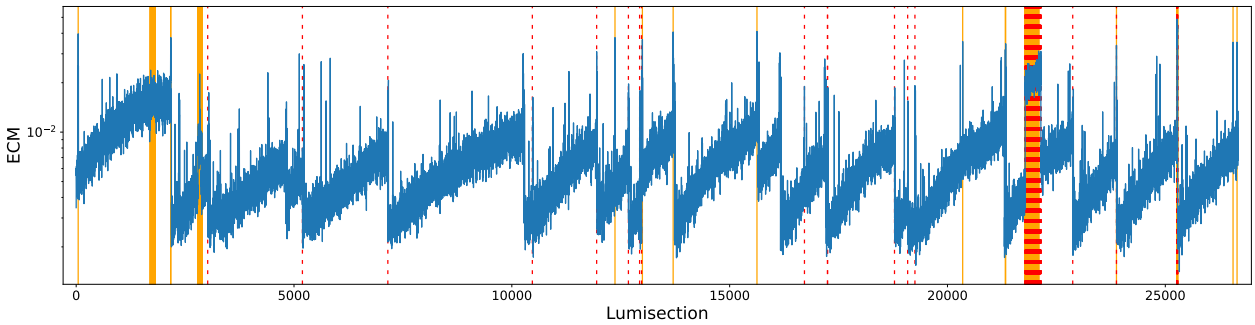


Figura A.10: ECM y etiquetas asignadas de η en la era C.

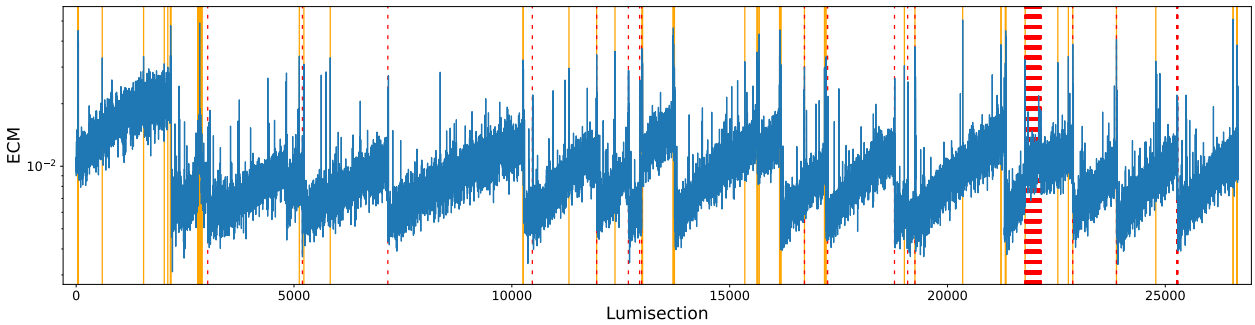


Figura A.11: ECM y etiquetas asignadas de φ en la era C.

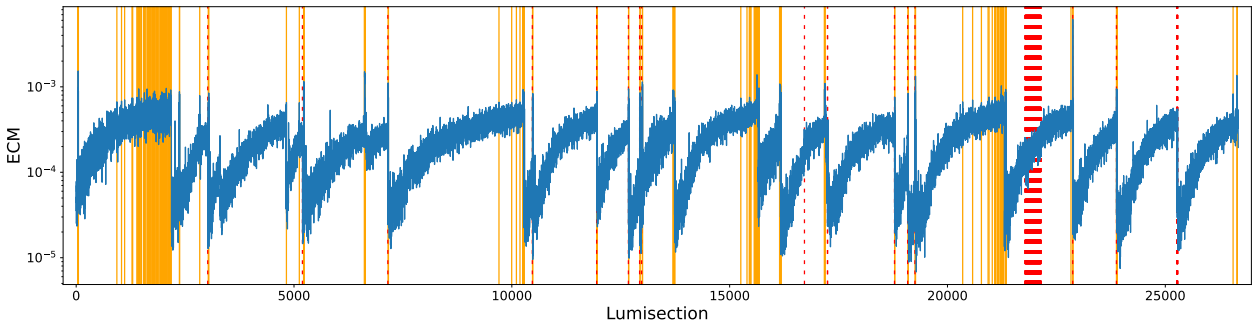


Figura A.12: ECM y etiquetas asignadas de χ^2/g.l. en la era C.

Evaluación en la era D. Entrenamiento en A, B y C.

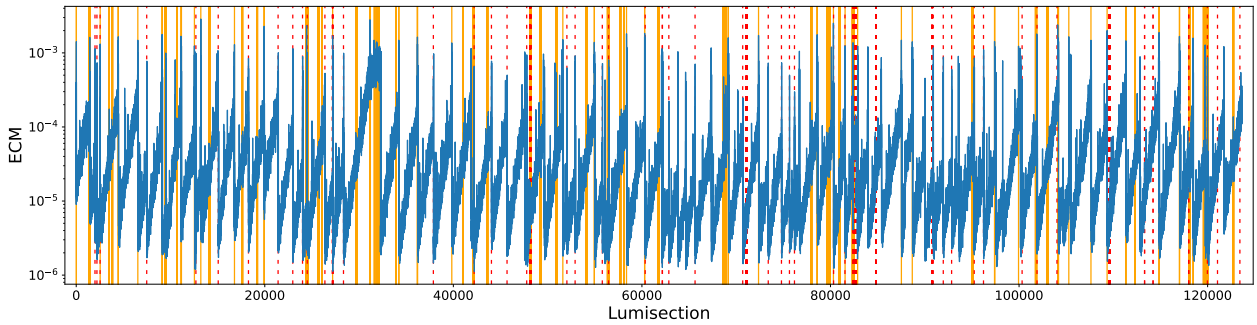


Figura A.13: ECM y etiquetas asignadas de p_t en la era D.

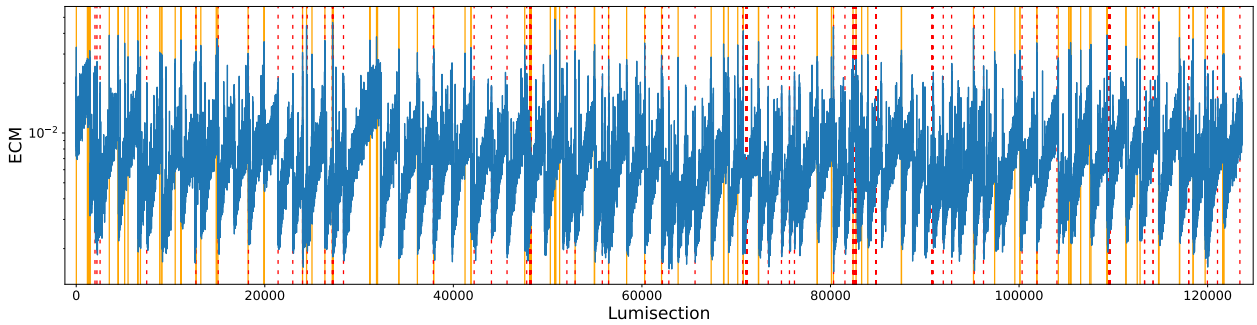


Figura A.14: ECM y etiquetas asignadas de η en la era D.

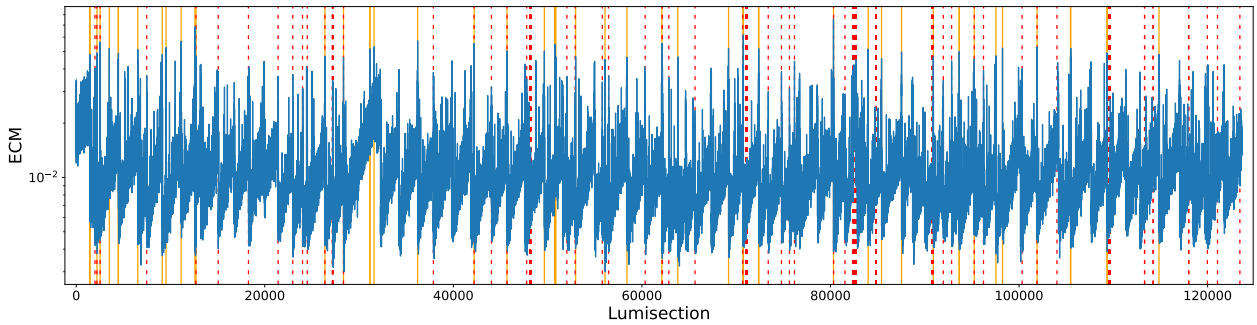


Figura A.15: ECM y etiquetas asignadas de φ en la era D.

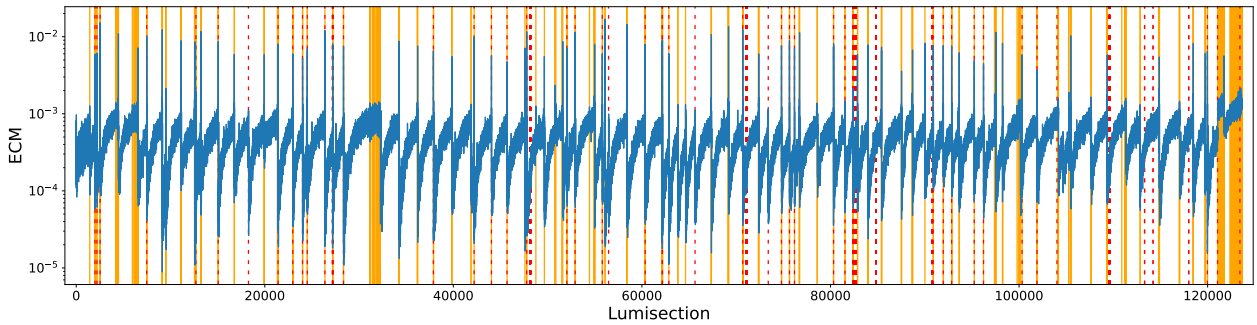


Figura A.16: ECM y etiquetas asignadas de χ^2/g.l. en la era D.

Anexo B: Código del trabajo

El análisis de los datos de muones detectados en CMS y el diseño del modelo de aprendizaje automático presentado en este trabajo han sido llevados a cabo en el lenguaje de programación *Python*. El desarrollo de este código ha sido, sin duda alguna, una de las partes más relevantes del mismo, llevando a realizar múltiples pruebas debido a la gran variedad de enfoques posibles a la hora de abordar el problema que nos hemos planteado.

El código empleado se encuentra disponible en GitHub y se puede acceder a él en el enlace:

`https://github.com/manueligal/tfg-fisica`

Los dos archivos desarrollados son *main.ipynb*, que contiene el núcleo del trabajo a través del análisis de los conjuntos de datos de los muones, la elaboración de gráficas y el entrenamiento y evaluación del modelo, y *funciones.ipynb*, que contiene todas las funciones de elaboración propia necesarias para el correcto funcionamiento del archivo principal. Este último se ha separado con el fin de aligerar la carga de código del primero y facilitar su manejo.

Índice alfabético