Estado del Arte de Machine Learning y su Aplicación en el Experimento LHCb

Eyvilin Velasquez Sanchez

Asesora: Phd. Melissa María Cruz Torres

Maestría en Matemáticas con Orientación es Estadística Matemática, Universidad Nacional Autónoma de Honduras

10 de septiembre de 2022

Resumen: El machine learning (ML) es una rama de la inteligencia artificial que tiene como objetivo desarrollar algoritmos que permitan a la maquina aprender de la experiencia. El presente estudio tiene como objetivo dar una reseña histórica del Machine Learning o aprendizaje automático para descubrir como empezó todo y como evoluciono superando cada reto para convertirse en lo que es hoy en día. También, hablaremos sobre su importante aplicación en el Experimento LHCb siendo esto una pieza clave para la correcta clasificación de los datos. La determinación de un adecuado criterio de selección es fundamental para todo análisis en donde se extraen medidas Físicas a partir de grandes cantidades de datos. La correcta selección de eventos representa un desafió, pues, usualmente no únicamente están presentes eventos de señal sino que también eventos de ruido de distintas fuentes.

Palabras clave: Experimento Large Hadron Colieder beauty (LHCb), Machine Learning (ML), Decaimientos del mesón B.

1. Machine Learning: ¿Qué es y como funciona?

El termino de machine learning (ML) nació a principios de los años 50, no es fácil encontrar una definición, sin embargo, diremos que machine learning es una sub-disciplina de la inteligencia artificial cuyos algoritmos tienen la capacidad de aprender y tienen la capacidad de resolver problemas sin ser programados explícitamente para resolver esos problemas, ellos aprenden de los datos, logran inferir de los datos reglas para solucionar el problema.

También podemos ver a machine learning como la inferencia estadística llevada al computador, hay una mezcla entre dos campos de la ciencia de datos (ciencias de la computación y estadística) que es lo que da lugar a machine learning. Y se trata de encontrar el modelo que mejor describa a los datos.

En ML hay dos paradigmas más utilizados dependiendo de la naturaleza del aprendizaje:

1. Aprendizaje supervisado
2. Aprendizaje no supervisado

En el aprendizaje supervisado al programa se le proporcionan los datos de entrada (o inputs) de los que tiene que aprender asociados con las respuestas (o outputs) que para ellos se esperan del programa. A este conjunto de datos se le llama conjunto de entrenamiento (training). El programa se encarga de alterar sus parámetros para que la respuesta sea óptima. Dependiendo de si la respuesta es bien un número, bien una etiqueta (label) o clase, será un problema de **regresión** o de **clasificación**. En el caso del aprendizaje no supervisado al programa no se le proporcionan las respuestas deseadas, sino solamente los datos de entrada, y él debe encontrar algún tipo de estructura en ellos, detectando por ejemplo, características desconocidas. Un ejemplo de ellos son los problemas de agrupación o clustering. [3]

El objetivo del aprendizaje supervisado es hacer inferencia estadística de la distribución que generan los datos y predecir el output "y_{new}"dada una nuevo input "x_{new}".

En ML tenemos dos categorías de datos diferentes, los *datos de entrenamiento* con los que el algoritmo va a aprender y hacer el modelo y los *datos de evaluación o test* con los cuales se va a evaluar el modelo para

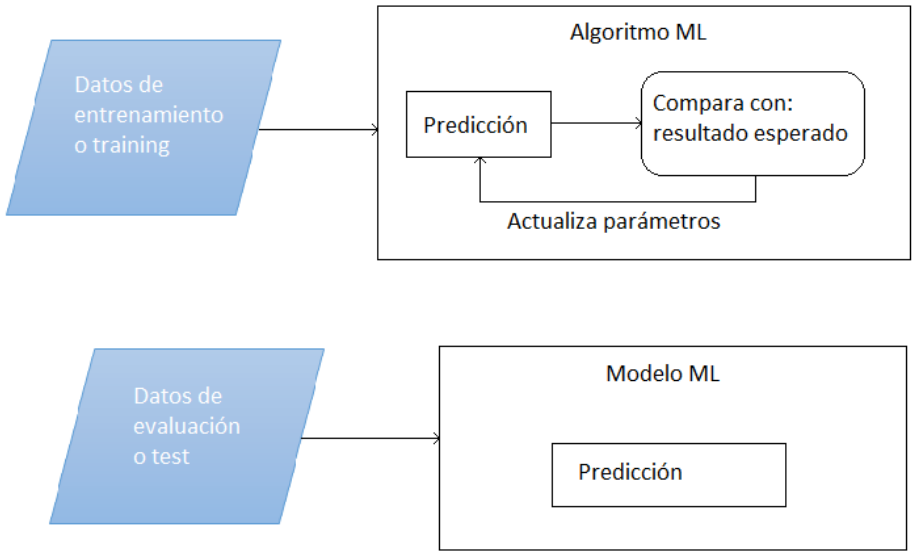

Figura 1: Algoritmo de machine learning

ver si aprendió y no hizo un sobre ajuste “overfitting”.

El funcionamiento del entrenamiento en los algoritmos de machine learning se puede visualizar en la Figura 1. El objetivo es hacer una predicción y lo compara con los resultados esperados, dependiendo que tan bien le haya ido, el algoritmo va a hacer una retroalimentación actualizando los parámetros, esto es un ciclo. Luego que el algoritmo esta entrenado tenemos un modelo de ML, al cual le pasamos los datos de evaluación para ver si logra predecir bien sin caer en un sobre ajuste.

Las técnicas de machine learning constan aproximadamente de tres pasos: *entrenamiento o training* en las muestras de simulación para aprender a discriminar entre señal y ruido, *prueba o testing* en muestras de simulación independientes para evaluar los resultados y la evaluación de la salida o predicción en muestras de datos.

2. Historia del Machine Learning

En esta sección se detalla la historia del machine learning, algunas personas pueden pensar que es algo muy reciente sin embargo, la historia del machine learning se remonta muchos años atrás. Para conocer los inicios de machine learning y como las personas en busca de la necesidad de analizar los datos, obtener información de ellos y automatizar este procedimiento dieron origen al mismo. Sin embargo, no es posible decir cuando se inventó el aprendizaje automático o quién lo inventó, podemos decir que es el resultado de la combinación del trabajo de muchas personas dedicadas a las investigaciones, inventos y algoritmos. En este viaje se darán algunos momentos claves para el surgimiento de lo que ahora es el machine learning.

Se comenzará este viaje en el año de 1642, Blaise Pascal creo la calculadora mecánica con el objetivo de automatizar el procesamiento de datos. Sorprendentemente en 1801, Joseph Marie Jacquard fue uno de los pioneros en automatizar el proceso del tejido, el cual utilizaba tarjetas metálicas con agujeros para organizar los hilos. Este es considerado el primer dispositivo de almacenamiento de datos.

Con la introducción de la lógica booleana en 1847 y la maquina de Hollerith en 1890, el cual es un sistema electromecánico que se encarga de cálculos estadísticos. Esta maquina se utilizó en el censo de los Estados Unidos en 1890, para ayudar a procesar los datos, esto dio origen a la industria del procesamiento de datos.

En el año de 1943 llega un momento clave en la historia del machine learning con la publicación del artículo científico “A Logical Calculus of the

Ideas Immanent in Nervous Activity", publicado por el lógico Walter Pitts y el neurocientífico Warren McCulloch, ellos sin saberlo estaban asentando las bases de lo que hoy en día son las redes neuronales artificiales y el aprendizaje profundo. Para varias personas este artículo fue el comienzo de formular matemáticamente los procesos de pensamiento y la toma de decisiones en la cognición humana, es decir, como se comporta una neurona y su participación en la capacidad de computar y procesar la información.

En 1950, Alan Turing un informático, criptoanalista, matemático y biólogo teórico inglés creó la "Prueba de Turing" para determinar si una computadora tiene inteligencia real, es decir si una computadora piensa o no como humano. En ese mismo año, Turing publicó un artículo titulado «Computing Machinery and Intelligence» en el que describía la prueba. La prueba consistía en que una computadora debía ser capaz de engañar a un humano haciéndole creer que también es humano. Turing es considerado un pionero del machine learning en los años 40 y 50.

En el año de 1952, se crea el primer programa de aprendizaje por computadora desarrollado en IBM, diseñado por Arthur Samuel. El programa era el juego de damas que se escribió sobre el algoritmo de poda alfa-beta un algoritmo de búsqueda que disminuye la cantidad de nodos evaluados por el algoritmo minimax en los árboles de búsqueda.

La Figura 2 es de 1956, donde por primera vez se mostró al publico el juego de damas. Donde Robert Nealy (Autodenominado maestro de las damas) jugó con el programa en una computadora de IBM, la computadora ganó.

Lo sorprendente del programa era que cuanto mas jugaba, mas aprendía, es decir, mejoraba su rendimiento en el juego aprendiendo nuevas estrategias. Es un modo de lo que conocemos hoy en día como aprendizaje supervisado. La creación de este programa fue algo muy innovador para la época y sentó las bases para que las máquinas hicieran otras tareas inteligentes mejor que los humanos. Siendo así, que las personas comenzaban a cuestionarse si las computadoras lograran superar a los humanos en inteligencia.

Ahora bien, en 1957 el inventor Frank Rosenblatt, en el laboratorio Aeronáutico de Cornell, diseñó la

Figura 2: El 24 de febrero de 1956 un programa de ordenador diseñado por Arthur Samuel, investigador de IBM, derrotó a una persona jugando a las damas. (Fuente: IBM).

primera máquina capaz de generar un pensamiento original, es considerada la primera red neuronal para computadoras "el perceptrón", este se planeo inicialmente como una maquina no como un programa. El perceptrón simulaba los procesos de pensamiento del cerebro humano y fue categorizada como la primera neurocomputadora exitosa. Aunque el perceptrón parecía prometedor, se tenían expectativas altas, no podía reconocer muchos tipos de patrones visuales como los rostros, lo cual mantuvo paralizada paralizaba la investigación de redes neuronales por muchos años.

Un gran avance en la historia del machine learning se da cuando las computadoras adquieren la capacidad de reconocer patrones. En 1967 se escribió el algoritmo del "vecino más cercano". Uno de los algoritmos mas importantes que resolvió el problema del vendedor ambulante de encontrar la ruta mas eficiente, el cual consiste en que un vendedor comienza en una ciudad aleatoria y visita las ciudades vecinas repetidamente hasta que todas han sido visitadas. En el algoritmo cada vez que el programa recibía un nuevo objeto, hacía una comparación con los elementos existentes y lo clasificaba como el elemento mas similar dentro de los elementos guardados, es decir, el vecino mas cercano. Este avance del reconocimiento de patrones sentaron las bases de lo que ahora conocemos como la Inteligencia Artificial (IA).

Fue hasta en 1979 que los estudiantes de la Universidad de Stanford inventaron el primer vehículo

autónomo el “carro de Stanford“, que era capaz de maniobrar obstáculos en una habitación. Fue un gran proyecto que duro varios años entre 1960 y 1980.

En 1981 Gerald Dejong publicó un artículo siendo el pionero en el concepto de aprendizaje basado en explicaciones (EBL), el mismo sentó las bases de lo que ahora conocemos como aprendizaje supervisado.

En la década de 1990 los científicos dan un giro al trabajo del aprendizaje automático, el cual cambia de un enfoque basado en el conocimiento a un enfoque basado en datos. Los científicos empiezan a crear programas que sean capaces de analizar grandes cantidades de datos y puedan aprender de sus resultados. Este fue el inicio de la minería de datos, las aplicaciones web, el aprendizaje de textos y de idiomas. El éxito de tener grandes cantidades de datos fue el resultado del crecimiento del Internet.

El término de ML tomó más fuerza cuando los científicos lograron desarrollar programas que fueran capaces de aprender por sí mismos. Estos programas incluyen los algoritmos de máquinas de vectores de soporte, árboles de decisión y bosques aleatorios. El concepto de Boosting o fuerza se implemento por primera vez en un artículo publicado en 1990 titulado “La fuerza de la capacidad del aprendizaje débil“, de Robert Schapire. El Boosting ayudó a la evolución del machine learning, dado que nos brinda la ventaja de que reduce el sesgo durante el aprendizaje supervisado, la idea es transformar alumnos débiles en uno solo alumno fuerte.

En 1997, sucedió un hecho que sobresalto a las personas al ver que Deep Blue de IBM, venció al campeón mundial de ajedrez, basándose en el análisis de partidas previas.

En 2006, Geoffrey Hinton, acuña el término de “Deep Learning“ para explicar los nuevos algoritmos que llevan a las computadoras a poder distinguir objetos y texto en imágenes y vídeos. Los populares algoritmos de reconocimiento facial de esa época fueron evaluados por un programa del Instituto Nacional de Estándares y Tecnología, los hallazgos indicaron que estos algoritmos eran diez veces mas potentes que los creados en años anteriores. Incluso algunos de los algoritmos lograron superar a los humanos en el reconocimiento de rostros y pudieron a gemelos idénticos.

Luego cuatro años mas tarde, Microsoft lanzó el dispositivo de entrada de detección de movimiento Kinect para su consola de juegos Xbox 360, que podía rastrear 20 rasgos humanos diferentes a una velocidad de 30 veces por segundo. Al siguiente año, Watson de IBM venció a sus competidores humanos en Jeopardy.

También en el 2010, se dio el lanzamiento de Kaggle por Anthony Goldbloom y Ben Hamner, la idea original fue de una plataforma para competencias de aprendizaje automático.

En 2011, el proyecto de Google Brain fue lanzado, al año siguiente el equipo de Google Brain desarrollo un algoritmo de ML en especifico creo una red neuronal que es capaz de aprender a reconocer gatos a través de vídeos de YouTube.

En 2014, Facebook desarrolló DeppFace, un algoritmo de ML capaz de reconocer o identificar personas en fotografías con la misma precisión que los humanos. En el mismo ano Google presento Sibyl, su proyecto de ML a gran escala para recomendaciones predictivas de usuarios.

En 2015, Amazon lazó su propia plataforma de ML. Por otro lado, Microsoft crea el kit de herramientas de ML distribuido que ayudó hacer una distribución mas eficiente de problemas de ML en varias computadoras. En ese mismo año Stephen Hawking, Elon Musk y Steve Wozniak y mas de 3000 investigadores de Inteligencia Artificial y robótica, firmaron una carta abierta advirtiendo sobre el peligro de las armas autónomas que seleccionan y atacan objetivos sin intervención humana.

En 2016, sucedió una de las victorias mas populares del ML gracias al algoritmo AlphaGo desarrollado por Google Deepmind, que venció al campeón mundial en el juego de mesa chino “Go“, el cual es considerado el juego de mesa mas complejo del mundo siendo muchas veces mas difícil que el ajedrez. El algoritmo AlphaGo logro ganar cinco juegos de cinco en la competencia del juego de mesa Go. AlphaGo y sus sucesores vencieron a varios campeones de Go, Ajedrez y Shogi. AlphaGo utiliza técnicas de aprendizaje automático y búsqueda de árboles.

Waymo comenzó a probar sus minivans autónomas

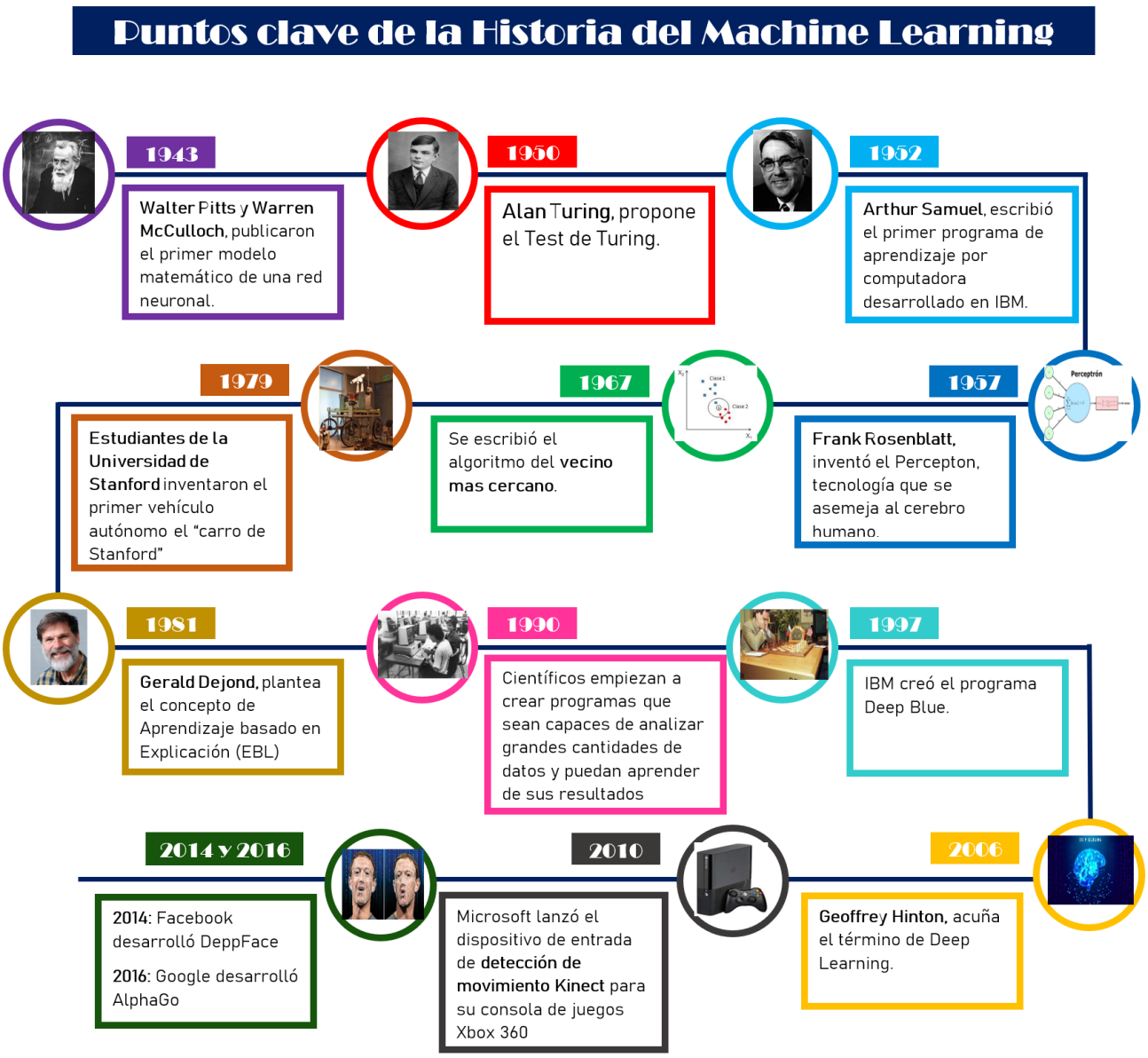

Figura 3: Breve recorrido por los puntos claves de la historia de Machine Learning

en los Estados Unidos en el 2017, con conductores de respaldo solo en la parte trasera del auto. Más tarde ese mismo año introducen taxis completamente autónomos en la ciudad de Phoenix. Ahora bien, en el 2018 Deepmind con el algoritmo AlphaFold fue un éxito por su capacidad para predecir la estructura de las proteínas.

En el 2018, Google Deepmind comenzó a desarrollar sistemas que pueden competir en una variedad de juegos. Les permite probar las capacidades de los sistemas ya que varios juegos modifican sus comportamientos. El objetivo de la gamificación es permitir que el sistema aprenda a adquirir inteligencia y comportamiento similares a los humanos y con ello aprovechar las capacidades del ML.

En 2020, llego la pandemia algo totalmente inesperado para todos; sin embargo la IA creo lo que se considera hoy en dia el modelo de lenguaje mas grande y avanzado del mundo "GPT-3", el cual es un innovador algoritmo de procesamiento de lenguaje natural. El GPT-3, tiene la capacidad de generar texto similar al humano cuando se le solicita, utiliza 175 000 millones de parámetros y la supercomputadora de IA de Microsoft Azure para la capacitación. Para mas deta-

lles de la historia del ML ver [4], [8], [6].

En la Figura 3 se brindan los puntos claves de la historia del ML.

3. El Futuro del Machine Learning

Es difícil imaginarnos un futuro donde no hagamos uso del ML en nuestras vidas cotidianas. Dado que el ML esta detrás de muchas tecnologías y de empresas lideres como Google, Amazon, Facebook, Netflix, Tesla y muchas mas utilizan el ML de manera eficiente. El ML juega un papel importante en las empresas, ya que permite a sus dueños comprender el comportamiento de sus clientes, esto ayuda a la toma de decisiones y a la vez a entender el funcionamiento del negocio. El tamaño del mercado global de ML está valorado en $21,17 mil millones en 2022 y se espera que alcance $209,91 mil millones para 2029, según Fortune Business Insights.

Pensar en el futuro del aprendizaje automático es sumamente emocionante. En la actualidad, casi en todos los ámbitos están presentes aplicaciones que funcionan con ML. Por ejemplo, en el marketing digital, educación, salud, los motores de búsqueda entre muchos mas. Los campos de la visión por computadora y el procesamiento del lenguaje natural (NLP) están logrando avances que nadie se imaginaba. Todos estos avances los vemos en nuestro día a día, en nuestros teléfonos inteligentes con el reconocimiento facial, software de traducción de idiomas, automóviles autónomos, por mencionar algunos. Lo que veíamos en las películas de ciencia ficción se están volviendo realidad.

Estas son algunas predicciones sobre el desarrollo del ML en los anos futuros.

- **Mejoras en los algoritmos de aprendizaje no supervisado**
  Esto ayudaría a hacer predicciones a partir de datos no etiquetados. Esto ficción se volverá cada vez mas importante, ya que da la facilidad que los algoritmos descubran patrones ocultos o grupos que ayuden a comprender de la manera mas eficiente un problema.

- **Computación cuántica**
  La computación cuántica podría transformar el futuro de ML, las computadoras cuánticas conducen a un procesamiento de datos más rápido, lo que mejora la capacidad del algoritmo para analizar y extraer información significativa de los conjuntos de datos obteniendo conocimientos mas profundos.

  Hasta el momento, no existe un modelo de ML cuántico disponible comercialmente. Sin embargo, el CERN el mayor laboratorio de Física de Partículas esta incursionando en este campo. Se hablara mas de ello en la siguiente sección.

- **Machine Learning Automatizado o AutoML**
  Se trata de aplicar algoritmos de ML a tareas de la vida real. El autoML se puede automatizar en los siguientes procesos: Preprocesamiento de datos, selección de funciones, optimización de hiperparametros y selección de algoritmos, entre otros.

- **Industrias a tener en cuanta para la aplicación del ML**
  Industria manufacturera, el ML puede mejorar, la calidad del producto, reducir costos, marketing hasta las ventas y el mantenimiento de las maquinas.
  Industria farmacéutica y de atención médica, la industria de la salud genera una masiva cantidad de datos que ayudan a optimizar los procesos administrativos y tratar enfermedades infecciosas. Cada vez las enfermedades están evolucionando y es necesario identificarlas y poderlas diagnosticar antes.
  Industria Automotriz, el ML no solo nos lleva a conducir autos autónomos, sino también mejorar desde la investigación hasta el diseño y la fabricación.

4. Aplicación de ML en el Experimento LHCb

El volumen de datos recabados por los experimentos del LHC ha crecido rápidamente, muestra de ello nos la da el Run II que tuvo lugar en los años 2015 - 2018 que logro doblar la cantidad de eventos del Run I (2011 - 2012), esto hace que los retos de análisis de datos y computacionales sean grandes. Para mas informacion consultar [7].

El experimento LHCb utiliza el ML para distintos fines, por ejemplo, en la reconstrucción de trazas lo cual requiere inferir las trayectorias de las partículas cargadas, no es una tarea fácil. Se utiliza algoritmos

de ML como GNNs (graph neural networks).

Dado que las colisiones ocurren cada 25 μs, se cuenta con una increíble cantidad de datos. Esto implica que se pierdan eventos significativos. Es por este motivo que el detector LHCb tiene que tener un sistema eficiente que nos permita seleccionar únicamente datos de interés, es decir, un primer gran filtro dado que la capacidad de almacenamiento costoso y es finito. Ese sistema de selección de eventos del experimento LHCb se llama *Trigger*, el cual tiene dos niveles el trigger de nivel-0 (L0) en software y el trigger de alto nivel (HLT) en hardware, este a su vez se divide en dos subniveles (HLT1 y HLT2). La primera etapa HLT1 se hace una reconstrucción parcial de eventos, esta busca partículas con alto P_T las cuales no se originan en el vértice primario. HLT2 se realiza la reconstrucción total de los eventos y se hace una selección de datos exclusiva e inclusiva. En esta etapa se utilizan algoritmos más rigurosos como redes neuronales y algoritmos basados en la topología de los decaimientos de dos, tres o cuatro cuerpos. Para mas detalles ver [2].

Uno de los pilares fundamentales en el éxito del estudio de procesos de decaimiento es la correcta identificación de las partículas en el estado final (π y K) y entonces de la correcta identificación de los diferentes canales de decaimiento. Las técnicas de Machine Learning son también aplicadas en la reducción de ruido blanco o combinatorial. Este ruido y otros contaminan la región de señal quedando abajo o alrededor de ella, lo que implica que no sea una distribución Gaussiana como es esperado. La identificación y aislamiento de ese ruido no es fácil, porque han perdido una partícula o se les ha asignado una identidad que no es la de ellas. Esto se convierte en una pieza clave para la obtención de una muestra limpia de señal, lo cual es fundamental para extraer posteriormente medidas físicas. Técnicas multivariadas como redes neuronales, arboles de decisión y bosques aleatorios son muy utilizadas para la clasificación de partículas.

4.1. Grupo de trabajo de aprendizaje automático interexperimental del LHC (IML)

La enorme cantidad de datos representa un gran desafió que requieren el uso de los algoritmos de ML. La implementación puede ir desde escalas pequeñas que requieren poco tiempo hasta escalas grandes que requieran mas tiempo. El grupo de trabajo de aprendizaje automático interexperimental del LHC reúne a científicos del LHC, se brindan capacitaciones, se presentan los problemas y se plantean soluciones interexperimentales, se organizan eventos de difusión y entrenamiento para la comunidad experimental, se organizan reuniones mensuales sobre diversos temas centrándose siempre en determinada técnica de ML. Cada experimento del LHC esta representado por un coordinador. Uno de los objetivos de este grupo es facilitar la comunicación entre los experimentos. Puede encontrar mas información en [1]

4.2. Quantum Machine Learning prometedor en el experimento LHCb del CERN

La colaboración LHCb en el CERN, informo sobre un primer trabajo que esta relacionado con la computación cuántica, esto es algo que se menciono en la sección anterior. Esto se dio a conocer en una publicacion en el Journal of High Energy Physics, el articulo describe la aplicación de Quantum Machine Learning a la identificación de Jets que se originan a partir de quarks de belleza o antiquarks.

El trabajo fue realizado en la Universidad de Padua en Italia por el Grupo de Análisis y Procesamiento de Datos LHCb, en el que participa la Universidad de Maastricht. El coautor Davide Nicotra se unió desde entonces a la Universidad de Maastricht como estudiante de doctorado, donde continuará trabajando en aplicaciones de computación cuántica para los desafíos de seguimiento de partículas LHCb. Para mas detalles ver [9].

5. Conclusión

Se ha hecho un recorrido sobre como surgió y evoluciono el ML, logrando superar distintos tipos de retos para convertirse en lo que es en la actualidad, teniendo un sin numero de aplicaciones en todos los ámbitos. El futuro de ML nos dice que abrirá múltiples oportunidades para las empresas, el diagnostico

temprano de las enfermedades y la optimización de varios procesos en distintas industrias. A medida que las tecnologías continúen desarrollándose, los algoritmos de ML se pueden usar de una manera mas efectiva.

La implementación de las técnicas de machine learning son cada vez mas importantes en los experimentos del LHC, incorporar técnicas como Deep Learning se hacen mas necesarias. Se considera que es muy importante el trabajo interdisciplinar, para trabajar con la gran cantidad de datos que se tienen, logrando clasificar correctamente los eventos e interpretarlos de manera adecuada.

Referencias

[1] IML: Inter-Experimental LHC Machine Learning Working Group (2018). Iml, "who we are. ginebra, suiza,". https://iml.web.cern.ch/ Revisado el 20 de marzo, 2020.

[2] Denis Derkach. Machine-learning-based global particle-identification algorithms at the lhcb experiment. *Journal of Physics: Conference Series*, 2018.

[3] Mario Señas Gómez. Simulaciones realistas de colisiones protón-protón en el lhc usando una red neuronal convolucional extractora de correlaciones locales. 1, 8 2018.

[4] Nikita A. Kazeev. Machine learning for particle identification in the lhcb detector. *NATIONAL RESEARCH UNIVERSITY HIGHER SCHOOL OF ECONOMICS*, 2020.

[5] W. H. McCulloch, W. S. y Pitts. A logical calculus of the immanent in nervous activity. 1943.

[6] Victor Nalda. Machine learning: Los orígenes y la evolución. 2020.

[7] Cristina Oropeza-Barrera. Uso de técnicas de machine learning en el experimento cms. 2020.

[8] HASAN SELMAN. The history of machine learning – dates back to the 17th century. *https://dataconomy.com/2022/04/the-history-of-machine-learning*, 2022.

[9] Arthur Pesah Maria Schuld Koji Terashi Sofia Vallecorsa Jean-Roch Vlimant. Wen Guan, Gabriel Perdue. Quantum machine learning in high energy physics. *arXiv:2005.08582v2*, 2020.