# Aplicación de Algoritmos de Machine Learning para la Identificación de Partículas en el Experimento LHCb

Velasquez, Eyvilin (Maestría en Matemáticas con Orientación en Estadística, UNAH); eyvilin.velasquez@unah.edu.hn [1];

Cruz, Melissa (Departamento de Altas Energías, Cosmólogos y Radiaciones, UNAH); melissa.cruz@unah.edu.hn [2];

**Resumen:**

El experimento LHCb ubicado en el Centro Europeo para la Investigación Nuclear (CERN) tiene como objetivo el estudio del fenómeno de violación de la simetría de Carga - Paridad o violación CP y de decaimientos raros en hadrones con contenidos de quarks *b* y c. La violación CP es uno de los ingredientes claves para entender la diferencia de materia y antimateria en el universo. El CERN es uno de los mayores y más prestigiosos centros de investigación en el mundo ubicado en la frontera entre Francia y Suiza.

La presente investigación tiene como objetivo la aplicación de algoritmos de machine learning basados en técnicas de análisis multivariante para la reducción del ruido en los canales $B^+ \to h^+h'^+h''^-$ particularmente los canales $B^+ \to \pi^+\pi^+\pi^-$ y $B^+ \to K^+K^-\pi^+$ que son los estudiados en esta investigación. Se utilizó parte de la muestra de datos recolectadas por el experimento LHCb en el Run 2, los cuales comprenden los años 2015 y 2016. 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.

En los canales $B^+ \to \pi^+\pi^+\pi^-$ y $B^+ \to K^+K^-\pi^+$ el ruido denominado como combinatorial es reducido considerablemente utilizando Boosted Decision Trees,

[1] Filiación Institucional (Unidad Académica, Universidad); Correo Electrónico Autor 1
[2] Filiación Institucional (Unidad Académica, Universidad); Correo Electrónico Autor 2

obteniendo una muestra más limpia de señal. En una segunda etapa el ruido de mis-identificación es reducido logrando una mejora en la distribución propia de los datos. Esta investigación presenta en detalle los estudios realizados para la determinación de este criterio de selección a través de herramientas estadísticas y algoritmos de machine learning. La eficiencia de la señal es de 88.09% y 82.85% para $B^+ \to \pi^+\pi^+\pi^-$ y $B^+ \to K^+K^-\pi^+$, respectivamente.

**Palabras Clave:**

Análisis multivariante (MVA), Boosted Decision Trees (BDT), Machine Learning (ML), Large Hadron Collider beauty (LHCb), Decaimientos del mesón *B*.

**Introducción:**

El experimento LHCb ubicado en el Centro Europeo para la Investigación Nuclear (CERN, por sus siglas en inglés) tiene como objetivo el estudio del fenómeno de violación de la simetría de Carga - Paridad o violación CP y de decaimientos raros en hadrones con contenidos de quarks *b* y *c*. La violación CP es uno de los ingredientes claves para entender la diferencia de materia y antimateria en el universo. El CERN es uno de los mayores y más prestigiosos centros de investigación en el mundo ubicado en la frontera entre Francia y Suiza. Es de resaltar que el Gran Colisionador de Hadrones es el acelerador de partículas más grande y con mayor energía en el mundo.

Grandes cantidades de partículas son producidas durante las colisiones protones-protones en el Gran Colisionador de Hadrones (LHC), dichas partículas dejan registrado su paso a través del detector. Un tipo particular de decaimiento, de mucho interés, es de los mesones B en piones y kaones, el cual no posee en su estructura el quark encanto, como ser $B^+ \to \pi^+\pi^+\pi^-$, $B^+ \to K^+K^-\pi^+$, $B^+ \to K^+\pi^+\pi^-$ y $B^+ \to K^+K^+K^-$[3]. Estos decaimientos son altamente sensibles a violación CP y por tanto ofrecen un laboratorio para el estudio de nueva física. El análisis de datos conteniendo estos decaimientos,

[3] La carga conjugada se sobreentenderá en todo el texto

cantidades de la orden de TB, es de alta complejidad. Uno de los pilares fundamentales en el éxito del estudio de estos procesos es la correcta identificación de las partículas en el estado final ($\pi$ 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 la parte esencial de este estudio y es una pieza clave para la obtención de una muestra limpia de señal.

Este trabajo describe un análisis estadístico, se utilizaron datos recopilados por la Colaboración LHCb en el LHC durante el Run II, que tuvo lugar del año 2015 - 2018. De este dataset se usó únicamente los datos 2015 y 2016. El estudio tiene como objetivo la aplicación de algoritmos de machine learning y la identificación de partículas, para la reducción de ruido en los canales $B^+ \to h^+h'^+h''^-$[4]. Serán utilizado como primer paso los métodos de análisis multivariante de variables que sean buenas discriminantes entre señal y ruido. Se realizará un estudio de eficiencia de cortes en las variables de identificación, también son construidas figuras mérito, donde se maximiza la significancia estadística.

**Descripción del método:**

El software de análisis de datos es Root y Python. Se utilizó la siguiente estrategia de análisis, primero se realizó la selección de análisis multivariante para la reducción del ruido combinatorial donde se utilizó algoritmos de machine learning basados en Boosted Decision Tree, luego se hizo un proceso de optimización para mis-identificación de

[4] $B^+ \to h^+h'^+h''^-$ donde $hh'h''$ representa kaones y piones

partículas donde se construyen figuras de méritos. Finalmente se realiza la aplicación de los cortes de BDT y PID a la distribución de probabilidad de masa invariante.

**Conjunto de Datos y Simulación:**

En este estudio se utilizó parte de la muestra de datos recopilada por el LHCb en el Run 2, los cuales comprenden los años 2015 y 2016 a una energía del centro de masa de $\sqrt{s} = 13$ TeV y una luminosidad integrada de 1.9 $fb^{-1}$, la cantidad de datos del 2016 aproximadamente es el doble en comparación al 2015. También se utilizaron muestras de simulación de Monte Carlo (MC) del año 2015 y 2016 las cuales reproducen al máximo las condiciones reales de las colisiones, son muestras totalmente limpias sin ningún tipo de ruido, es decir, representan solamente la señal. Antes de utilizar estas muestras, las mismas pasan por el proceso de selección.

A las muestras de Monte Carlo se les aplica un corte basado en tablas de verdad (MCTruth o Trueid). Esto se debe a que el software de DaVinci, plataforma privada del experimento, no sabe identificar específicamente quien es la primera y segunda partícula, por lo tanto, acepta tantos piones como kaones, consecuencia de ello es que se hace un doble conteo en la muestra de señal y es por esta razón que se aplican dichos cortes, es un corte de preselección específico para cada canal de decaimiento.

Trabajar con la muestra de datos recopilada por el LHCb en el Run 2, los cuales comprenden los años 2015 y 2016, resultó un gran desafío ya que la conexión se realizaba a través de un servidor remoto del Centro Brasileiro de Pesquisas Físicas (CBPF) para acceder a los datos y realizar los análisis. Trabajar de esta forma en cierto sentido, tomó más tiempo de ejecución en los algoritmos implementados en este artículo.

**Variables de interés:**

Las variables de interés se basan en características topológicas de los mesones *B*. En general, para la selección se está interesado en variables que sean buenas

discriminantes entre señal y ruido. Esto se logra explorando las características topológicas de las variables. El mesón *B* que es un mesón pesado, se origina en el punto de interacción del vértice primario (PV), luego el mesón viaja a una distancia antes de decaer, esa distancia se denomina distancia de vuelo (FD), luego el mesón decae, justo donde decae se llama vértice secundario (SV) y decaen las tres partículas hijas.

Entre las variables más utilizadas para discriminar entre eventos de señal y eventos de ruido tenemos las siguientes:

- Masa invariante: La masa invariante del candidato *B* se calcula utilizando medidas de energía y momentos de partículas de estado final identificadas. La masa invariante se define así:

  $$M = \sqrt{E^2 - \vec{p}^2}$$

  Donde, $\vec{p} = \vec{p}_1 + \vec{p}_2 + \vec{p}_2$ es la suma de los 3-momentos hija y $E = E_1 + E_2 + E_3$ es la suma de sus respectivas energías.
- Vértice primario y secundario (PV-SV): El vértice primario (PV) de una partícula es donde se origina la partícula, es decir, es el punto de interacción de la colisión protón-protón. El vértice secundario (SV) es el punto donde la partícula decae.
- Momento transverso ($P_T$): Es el momento perpendicular a la línea que une el vértice primario y el vértice secundario, es decir, perpendicular a la línea de vuelo del mesón B
- Distancia de vuelo (FD): distancia de vuelo del vértice primario donde se origina las partículas al vértice secundario donde decaen las partículas en su estado final.
- Parámetro de impacto (IP): se define como la máxima aproximación de la trayectoria y el vértice primario. La variable IP nos ayuda a distinguir partículas que se originaron en el vértice primario y partículas producidas en el vértice secundario.
- FD $\mathcal{X}^2$: es el ajuste obtenido de la distancia de vuelo.
- IP $\mathcal{X}^2$: El chi-cuadrado del impacto de las madres y de las hijas

- Variables de identificación de partículas (PID): Los subdetectores de calorímetros, RICH y sistema de muones del detector LHCb, se combinan y nos brindan información sobre un conjunto único de variables PID. Estas variables PID son utilizadas como criterios de selección en los análisis para identificar piones, kaones, electrones y muones. Ahora bien, en este estudio se emplean las variables ProbNN para reducir el ruido de mis-identificación de $K - \pi$ y $\pi - K$.

**Selección:**

El objetivo de este estudio que representa un proceso físico de interés es el aislamiento de una señal. Esto se logra a través del proceso de selección, que es uno de los pasos más importantes en cualquier análisis de física de alta energía basado en muestras de datos. Estas muestras contienen diferentes tipos de ruido, por ejemplo, las muestras de datos que se utilizan en el análisis no son totalmente específicas de un canal de desintegración, es decir, que en la misma muestra de datos se pueden incluir muchos canales con el mismo estado final.

El objetivo de este proceso de selección es reducir al máximo los eventos que no provengan del decaimiento deseado, esto es a lo que llamamos ruido, pero siempre con el compromiso de mantener la mayor cantidad de eventos de señal deseados. Sin embargo, la mayoría de los métodos que se enfocan en eliminar el ruido, siempre afectan los eventos de señal. Es por este motivo, que el proceso de selección busca maximizar tres cantidades: la pureza, la eficiencia y la significancia estadística, discutiremos sobre ellas en la sección de optimización

**Proceso de Selección:**

La selección de datos para un decaimiento de tres cuerpos $B^+ \to h^+ h'^+ h''^-$ es un proceso que conlleva varios pasos.

El sistema de Trigger (A. A. Alves, 2018) es el primer gran filtro por el cual pasan los datos de la colisión protón - protón. Luego pasan una selección llamada de Stripping, aquí los datos almacenados se procesan aún más para separar los eventos por la física de interés, es decir, son un conjunto de cortes según la física que se desea estudiar, son cortes bien relajados basados en la cinemática del decaimiento para no afectar de ninguna forma nuestra señal, pero sí que permita diferenciar otros canales que no se tiene interés en analizar.

Ahora bien, la muestra de datos colectada por el detector contiene la familia de decaimientos $B^+ \to h^+h'^+h''^-$, siendo así, en la preselección se realiza un conjunto de cortes específicos para cada canal de decaimiento, con el objetivo de reducir el ruido físico. En este punto las muestras de datos no han tenido ningún requisito de identificación de partículas del estado final, lo cual conduce a una identificación errónea de piones y kaones llamado ruido de alimentación cruzada de otros canales $B^+ \to h^+h'^+h''^-$. Para reducir este tipo de ruido se aplican cortes flexibles a las variables PID de los dos canales de decaimiento, estas variables son ProbNNK que da el valor de probabilidad de que la partícula sea un Kaón (*K*) y ProbNNpi que da el valor de probabilidad de que la partícula sea un pión ($\pi$).

Los cortes mínimos de PID aplicados a las partículas de kaones y piones son, ProbNNK>0.1 y ProbNNpi>0.1, se deshace de los picos que son ruidos para las muestras. Esto permite tener una distribución de masa invariante más limpia como punto de inicio de análisis. Claro, más estudios necesitan ser hechos que es lo que se explica más adelante con la selección de identificación de partículas.

**Distribución del Mesón *B:***

Se estudiarán los canales $B^+ \to \pi^+\pi^+\pi^-$ y $B^+ \to K^+K^-\pi^+$, después de haber aplicado los tres primeros pasos del proceso de selección de eventos en la muestra de datos, se procederá a seleccionar un corte con la ayuda del análisis multivariante, dicho

corte ayudará a disminuir la cantidad de ruido blanco o ruido combinatorial. Cuando la distribución de la variable es Gaussiana nos da la confianza de que esa distribución pertenece a ese canal de decaimiento. Nuestro histograma inicial no es una Gaussiana limpia como esperábamos. A pesar de que los datos han pasado por varios filtros basados en la física y en la topología del decaimiento. Es un proceso que tiene muchas otras partículas o contaminación de otras partículas, es decir, tiene ruido.

**Tipos de ruido:**

Cuando la distribución no sigue una distribución Gaussiana se sospecha que esa distribución no pertenece a ese canal, lo cual es producto de diferentes tipos de ruido. Estos tipos de ruido logran pasar los criterios de selección de Trigger y Stripping e ingresan a la región de señal debido a varias razones. Para el decaimiento en estudio $B^+ \to h^+h'^+h''^-$ las principales fuentes dominantes de ruido son:

- Ruido Combinatorial: Es el resultado de la asociación aleatoria de tres trazos no relacionados que aparentan ser el decaimiento en estudio $B^+ \to h^+h'^+h''^-$. Esto significa que cuando se hizo la construcción un trazo que no tenía nada que ver con el decaimiento, otro trazo igual que no tenía nada que ver con el decaimiento y otro trazo en las mismas circunstancias coincidieron en un punto y se seleccionaron como si fuera el decaimiento en estudio, pero realmente ellos no nacieron de ese punto, entonces son simplemente trazos aleatorios. La ventaja de este ruido es que como son trazos totalmente descorrelacionados en general son ruidos bien portados.
- mis-identificación: Este ruido se debe a los decaimientos de tres cuerpos $B^+ \to h^+h'^+h''^-$ que se reconstruyen con partículas mal identificadas, es decir, este error se produce al identificar una partícula $\pi$ como un $K$ en su estado final. Este error de identificación de partículas conduce a un pico de masa más bajo o alto que la masa de *B*. Por ejemplo, cuando un $\pi$ se identifica como un $K$ esto conduce a un pico de masa más bajo dado que la masa del $\pi$ es menor que la masa del $K$ y esto

nos lleva a que la distribución del mesón B no tenga el comportamiento de una Gaussiana.

- Decaimientos parcialmente reconstruidos de cuatro cuerpos: Este tipo de ruido ocurre cuando tenemos un decaimiento en cuatro cuerpos y una de las partículas no se reconstruyó entonces este decaimiento de 4 cuerpos se confundió con uno de tres cuerpos y como no representa en sí mi decaimiento su masa va a quedar fuera de lo que se denomina mi región de señal. Este tipo de ruido se modela en el ajuste de masa invariante *B*.

Todos estos tipos de ruido pueden quedar abajo de la región de señal, dentro o alrededor de ella, como primer paso en este estudio es disminuir al máximo el ruido combinatorial para ello solo se utilizará la región de la masa invariante arriba de 5400 MeV ($B_m > 5400$), la región debajo de la señal contiene señal más ruido, aquí predomina el ruido de mis-identificación, es por tal motivo que esa área es altamente desafiante. Ese será el último paso eliminar la mayor cantidad de eventos de ruido con el compromiso de mantener nuestra señal.

**Selección Offline:**

Esta selección offline involucra los últimos dos pasos del proceso de selección, la selección de análisis multivariante (MVA) y la selección de identificación de partículas (PID), son los que tienen mayor relevancia en este estudio, estos buscan eliminar las dos principales contribuciones de ruido que afectan la señal. El ruido combinatorial y el ruido debido otros canales hadrónicos.

- Selección Análisis Multivariante: El método que se utilizará para clasificar las partículas con el objetivo de reducir el ruido combinatorial es mediante técnicas análisis multivariante (MVA), en este estudio se utilizará el algoritmo de Boosted Decision Tree (BDT) para encontrar ese corte óptimo de BDT, esto se lograra

mediante la optimización de las denominadas figuras de mérito como serán definidas posteriormente.

- Selección de Identificación de Partículas (PID): Después de los criterios de selección por los que han pasado los datos, los cuales son Trigger, Stripping, preselección y selección de análisis multivariante (enfocado a reducir el ruido combinatorial). El ruido restante es debido a una identificación errónea de piones o kaones llamada Mis-identificación estos forman el de contaminación hadrónica en 3-cuerpos y el ruido de decaimientos parcialmente reconstruidos. Para eliminar estos tipos de ruido restante se utiliza el análisis de identificación de partículas (PID), el cual juega un papel clave en el entorno de selección. Esta selección PID al igual que la selección de MVA, busca eliminar el ruido manteniendo la mayor parte de la eficiencia de la señal. La selección de PID que se encontrará debe ser más fuerte que los cortes flexibles de PID ya aplicados en la preselección.

**Optimización:**

La optimización se logra encontrando el corte óptimo en la salida de BDT, este corte tiene la finalidad de reducir el ruido combinatorial, pero con el compromiso de mantener la mayor eficiencia de la señal.

El corte se encuentra agregando la variable BDT a las muestras y recorriendo todos los posibles cortes de BDT, así encontramos el corte óptimo que maximice la \figura de mérito (FoM). La figura de mérito es una cantidad que es función de una serie de parámetros, por ejemplo, la significancia estadística y el número de eventos, el gran objetivo será maximizar esta cantidad. A continuación, definiremos algunas cantidades que se pueden maximizar:

- Significancia Estadística:

  $Significancia = \frac{S_{MC}}{\sqrt{(S+B)_{datos}}}$, donde, $S_{MC}$ es el número de eventos tomados de las muestras de señal de Monte Carlo y $(S+B)_{datos}$ es el número de eventos en la región de señal definida por $|B_m - 5279| < 40\ MeV/c^2$, tomada de los datos reales

con los cortes de preselección. Esta definición es diferente a la significancia estadística en el contexto de pruebas de hipótesis.

- Pureza:

  $Pureza = \frac{S_{MC}}{(S+B)_{datos}}$, donde, $S_{MC}$ y $(S+B)_{datos}$ fueron definidos en la significancia estadística.

- Eficiencia del corte BDT:

  $Eficiencia = \frac{S_{corte}}{S_{Total}}$, donde, $S_{corte}$ es el número de eventos de señal que pasan el corte y $S_{Total}$ es el número total de eventos de señal antes que se realice algún corte. Estas eficiencias del corte se determinan mediante muestras de simulación de señal (MC). Podemos decir, que la eficiencia de BDT va a ir disminuyendo a medida aumente el valor del corte.

- Eficiencias PID: Las variables de identificación de partículas (PID) no están bien descritas en las simulaciones de Monte Carlo (MC), variando de las que se encuentran en los datos, por lo tanto, estas eficiencias se logran obtener por medio de un paquete llamado **PIDCalib** que significa “Calibración de Identificación de Partículas" implementado en ROOT. Este paquete es un conjunto de herramientas para ayudar a los analistas a calcular la eficiencia de los criterios de selección de identificación de partículas (PID).

Mediante estas eficiencias estimadas para muestras simuladas se encontrará el criterio de selección PID para las variables ProbNN, el cual es el último paso en el proceso de selección. Dichas eficiencias se logran obtener como peso para cada polaridad de imán y cada trayectoria del estado final, estos pesos se van agregando evento por evento en la región de aceptación. Los pasos para la implementación del paquete PIDCalib.

El proceso de selección busca optimizar la pureza, la eficiencia del corte y la significancia estadística, no obstante, suele ser bien complicado maximizar varias cantidades simultáneamente, por esta razón vamos a maximizar la **significancia**

**estadística** que nos brinda un equilibrio óptimo entre la pureza y la eficiencia del corte. La optimización de estas cantidades es una etapa crucial de este análisis.

## Machine Learning

El termino de machine learning (ML) nació a principios de los años 50, no es fácil encontrar una definición, sin embargo, diré que machine learning es una subdisciplina 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 (Bishop, 2006).

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 ver si aprendió y no hizo un sobre ajuste "overfitting". Hay dos paradigmas más utilizados en ML dependiendo de la naturaleza del aprendizaje: aprendizaje supervisado y aprendizaje no supervisado. En este estudio haremos uso del aprendizaje supervisado al cual se le proporcionan los datos de entrada (inputs) que se asocian con sus variables respuestas (output). Si la respuesta es un numero o una clase, será un problema de regresión o clasificación.

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

## Motivación para utilizar métodos multivariantes (MVA)

Tradicionalmente se hacían cortes rectangulares a cada una de las variables y se definía la clasificación como el conjunto de cortes utilizados en todas estas variables, se tiene la desventaja que esto se vuelve ineficiente y más aún si se tienen muchas variables

es más costoso computacionalmente. Se debe comprender bien lo que se está haciendo, porque el rechazo o el corte de eventos de background se está haciendo variable a variable, cuando se hace el corte en una variable cambia toda la distribución y se hace de nuevo el análisis para saber cuál es el corte óptimo en las variables restantes, se sigue el mismo procedimiento hasta hacer el corte en la última variable es por ello que es ineficiente. A partir del comportamiento físico de los procesos de señal y de ruido, se pueden postular valores justificados para el corte que disminuyan lo más posible la presencia de eventos de ruido, B, respecto a los sucesos de señal, S.

El objetivo es encontrar una región donde S sea mucho mayor que B, por ello es usual que las cantidades discriminantes tales como la significancia estadística, la pureza y la eficiencia, entre otras se utilicen. Entonces una región de corte será mejor que otra cuanto mayor sea el valor de estos discriminantes.

El problema es que una sola variable no da valores elevados de significancias, eficiencias y purezas, se realizan varios cortes secuenciales para definir la región de trabajo. A este tipo de análisis que involucran una única variable se dice que son mono-variante o univariante. Estas variables individuales no tienen un gran poder de separación y por ende, sus cortes no son muy óptimos. De hecho, podemos decir que es imposible lograr la significancia, eficiencia y pureza requeridas utilizando cortes en una sola variable y es por este motivo que son necesarios métodos que combinen muchas variables diferentes.

Por lo tanto, es por ello que se utilizan estos métodos multivariados para clasificación, porque combina un conjunto de variables, con buena separación entre señal y ruido en una sola variable que es el resultado que deseamos. Esto permite el rechazo de ruido mediante cortes en una sola "variable" o discriminante, generada específicamente para ese propósito. Se pueden hacer varias combinaciones y entrenar

varios métodos al mismo tiempo, aparte de eso no deforma el espacio de fase que es algo muy deseable para la física, en general tiene un mejor rendimiento.

Se desarrollaron muchos métodos diferentes con el objetivo de realizar la clasificación de eventos simultáneamente con múltiples variables. Existen varios métodos de MVA entre ellos están Discriminante lineal (LD), Redes Neuronales Artificiales, Boosted Desicion Tree (BDT), Maquinas de Vectores de Soporte, Estimador de verosimilitud proyectiva. En el presente trabajo utilizaremos BDT para la clasificación entre señal y ruido, es una de las técnicas más utilizados para este propósito en física de partículas. Estos métodos se organizaron en un paquete de C++ llamado Toolkit for Multivariate Analysis (TMVA) (K. Albertsson, 2020) que se implementa en el marco de ROOT y tiene aplicaciones específicamente en física de altas energías.

El CERN ha desarrollado ROOT, un entorno de trabajo, orientado a objetos para el análisis de datos a gran escala y que provee métodos estadísticos. Se utilizará ROOT y el paquete TMVA para nuestro análisis. Existen otros paquetes entre ellos están: Neural Bayes y scikit-learn que es Machine learning in Python.

**Boosted Decisión Tree (BDT)**

. Un árbol de decisión es una estructura binaria que consta de un nodo raíz y varios nodos de rama y hoja. El nodo raíz y cada nodo de rama están asociados con una pregunta binaria que involucra una de las variables discriminatorias. El árbol se atraviesa desde la raíz hasta una de las hojas y el camino se determina por la respuesta a la pregunta asociada con cada nodo posterior. Las hojas representan un evento ya sea como señal o como background.

De esta manera el espacio de las variables de discriminación se corta de forma no lineal en diferentes regiones rectangulares, que están etiquetadas por señal o background. El entrenamiento se detiene en un nodo tan pronto alcanza un límite inferior critico de eventos (nEventsMin) en ese nodo.

Las hojas determinan si el evento es señal o background y se etiquetan de acuerdo con su pureza $p = \frac{S}{S+B}$, Si $p > 0.5$ el evento se considera señal y si $p < 0.5$ el evento se considera background. La calidad de la separación está definida por la llamada función de impureza, algunos ejemplos son: Índice de Gini: $p(1-p)$, Entropia cruzada: $-pln(p) - (1-p)\ln(1-p)$, Error de Mis Clasificacion: $1 - \max(p, 1-p)$, Significancia estadística: $\frac{S}{\sqrt{S+B}}$.

Una deficiencia de los BDT es su inestabilidad con respecto a las fluctuaciones estadísticas en la muestra de entrenamiento. Este problema se supera utilizando un Boosting (refuerzo), la idea de este es combinar una colección de clasificadores débiles en uno fuerte. TMVA proporciona tres métodos de Boosting diferentes: Adative Boosting (AdaBoost), Gradient Boosting y Bagging Boosting. Esto hace que los árboles de decision sean más robustos.

En este estudio se utilizó el AdaBoost el cual según (Byron P. Roe, 2008, Febrero 2) se define de la siguiente manera: Si tenemos N eventos totales en la muestra, suponga que construimos k árboles, $T_k$.

Definimos el error del árbol $k$ así: $eer_k = \frac{\sum_{i=1}^{N} w_i I(y_i \neq T_k(x_i))}{\sum_{i=1}^{N} w_i}$, donde

$w_i$: El peso del i-ésimo evento

$x_i$: Conjunto de variables discriminantes para el i-ésimo evento

$y_i = 1$ si el i-ésimo evento es un evento de señal y $y_i = -1$ si el i-ésimo evento es un evento de ruido.

$T_k(x_i) = 1$ si el evento se encuentra en una hoja tipo señal en el árbol i-ésimo y $T_k(x_i) = -1$ si el evento se encuentra en una hoja tipo ruido en el árbol i-ésimo.

$I(y_i \neq T_k(x_i)) = 1\ si\ y_i \neq T_k(x_i)\ y\ 0\ si\ y_i = T_k(x_i)$, es decir es igual a 1 si la hoja en la que se encuentra el evento es del mismo tipo que este y 0 en otro caso

Lo que mide $err_k$ es la proporción de eventos mal clasificados. Mientras mayor sea $err_k$, el árbol obtenido disminuirá su calidad. Ahora definamos el peso del árbol como: $\alpha_k = \beta ln\left(\frac{1-err_k}{err_k}\right)$

Donde $\beta$ representa la intensidad del boost y es un parámetro adicional del BDT, $\beta = 1$ es el valor utilizado en el método AdaBoost estándar. Por último, cada uno de los pesos del evento se cambia de la siguiente forma: $w_i \rightarrow w_i exp^{\alpha_k I(y_i \neq T_k(x_i))}$

Y se normaliza $w_i \rightarrow \frac{w_i}{\sum_{i=1}^{N} N w_i}$ De forma que los eventos mal identificados ven su peso aumentado y los bien identificados lo ven disminuido. Adicionalmente, cuanto mejor sea el árbol (mayor valor de $\alpha$), más se modifican los pesos, con lo que se consigue variar más sensiblemente los siguientes resultados para diversificar el poder de separación del BDT.

Con todas estas definiciones estamos en condiciones de definir la variable discriminante que utiliza los datos de todos los árboles generados. Para cada uno de los eventos estudiados se define la variable BDT o respuesta del BDT: $BDT(x) = \sum_{m=1}^{N_T} \alpha_m T_m(x)$.

## Resultados

Se seleccionó las mejores variables discriminantes que dan el mayor poder de separación entre señal y ruido, que se utilizan como variables de entrada en el entrenamiento de BDT para producir una función, que por sí sola tiene mayor poder de separación que todas las variables utilizadas para entrenarla. El objetivo es realizar el corte en una sola variable generada con ese propósito que permita el rechazo de ruido. Se utilizó el paquete de TMVA implementado en ROOT que está diseñado para aplicaciones de física de altas energías, desarrollado en el CERN. Se verifica si las variables seleccionadas tienen una alta correlación, en este caso no hay correlaciones altas.

Hay muchas técnicas para producir el entrenamiento de análisis multivariante (MVA), para este análisis se comparó el desempeño de 7 diferentes técnicas, las cuales

son: Boosted decision tree (BDT), Multilayer Perceptron (MLP), Linear Discriminants (LD), likelihood y BDT con algunas modificaciones por ejemplo BDTPCA (BDT con variables transformadas en componentes principales) y BDTG (BDT con el boosting del gradiente).

Se ejecuto el entrenamiento de MVAs con el paquete TMVA con las variables que son mejores discriminantes entre señal y ruido. Para los datos de training (entrenamiento) tomamos el 70% y para los datos de test (prueba) se tomó el 30%. Es importante resaltar que los datos que utilizamos en el entrenamiento no pueden ser utilizados para el análisis por esta razón solo tomamos una porción de los datos reales para el entrenamiento.

Para ambos canales $B^+ \to \pi^+\pi^+\pi^-$ y $B^+ \to K^+K^-\pi^+$, se comprueba que el BDT es notablemente mejor que las demás técnicas utilizadas, teniendo un comportamiento similar al BDTPCA5100, BDTG y BDTPCA. Seleccionamos el BDT como mejor opción. Cuando comparamos el BDT con MLP, están totalmente superpuestos, pero el BDT tiene un mejor rendimiento con menos tiempo de cálculo para el entrenamiento.

El corte óptimo se determina mediante un estudio de optimización. Calculando los valores de la señal y el ruido para varios puntos de corte, se puede determinar el corte óptimo que nos de la mayor significancia estadística. Para el canal $B^+ \to \pi^+\pi^+\pi^-$ la mayor significancia de los datos se obtiene en BDT>0.044, logrando una eficiencia de la señal del 88.09% y una pureza del 84.8%. Para un 88.09% de la eficiencia de la señal se logra reducir en un 97% los eventos de ruido, lo cual es un excelente resultado para los análisis posteriores. Ahora bien, para el canal $B^+ \to K^+K^-\pi^+$ la mayor significancia de los datos se obtiene en BDT>0.06, logrando una eficiencia de la señal del 82.86%. Se logra alcanzar una pureza del 78.76%. Para un 82.86% de la eficiencia de la señal se logra reducir en un 98% los eventos de ruido combinatorial.

**Aplicación del corte óptimo de BDT**

Ahora se verá la aplicación del corte óptimo de BDT a la distribución de probabilidad de $B_m$ para el canal $B^+ \to \pi^+\pi^+\pi^-$ cómo se puede observar en la Figura 1. Se ha aplicado

el corte óptimo de BDT>0.044 para el canal $B^+ \to \pi^+\pi^+\pi^-$. Donde podemos apreciar un pico mayor de eventos de señal, confirmando que se logra una alta reducción del ruido combinatorial que es precisamente uno de los objetivos de este estudio. Un resultado similar ocurre para el canal $B^+ \to K^+K^-\pi^+$.

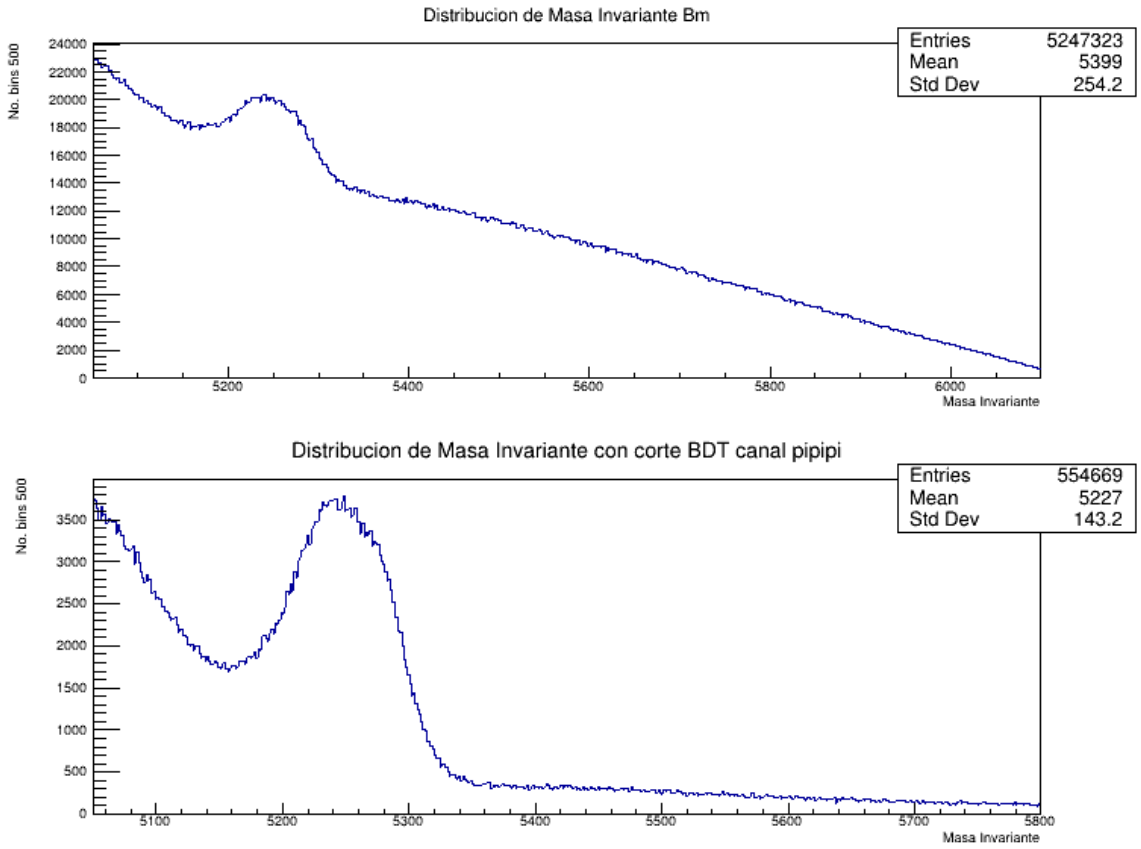


Figura 1: Distribución de $B_m$ (arriba) y Distribución de Probabilidad de $B_m$ con el corte de BDT (abajo), canal $B^+ \to \pi^+\pi^+\pi^-$.

**Selección de Identificación de Partículas (PID)**

Para tener un criterio de selección completa se realizó una optimización de las variables PID, estas variables son basadas en machine learning. Con el objetivo de obtener una señal utilizable para extraer una medida física. Este es el último paso para

reducir el resto del ruido en los cuales predomina el ruido de otros canales hadrónicos (identificación errónea de partículas) y decaimientos parcialmente reconstruidos.

Todo el análisis de las figuras de mérito se resume en el siguiente criterio de selección PID, que es el último paso del proceso de selección. Para el canal $B^+ \to \pi^+\pi^+\pi^-$, el criterio de selección PID es d1_ProbNNpi>0.521&d2_ProbNNpi>0.15& d3_ProbNNpi>0.191. Para el canal $B^+ \to K^+K^-\pi^+$, el criterio de selección PID es d1_ProbNNK>0.311&d2_ProbNNpi>0.266& d3_ProbNNK>0.422.

En las Figura 2, se muestra la aplicación del criterio de selección PID a la distribución de masa invariante $B_m$ para $B^+ \to K^+K^-\pi^+$. La distribución de masa invariante ya tiene aplicado la selección de análisis multivariante, es decir, ya tiene el corte óptimo de BDT.

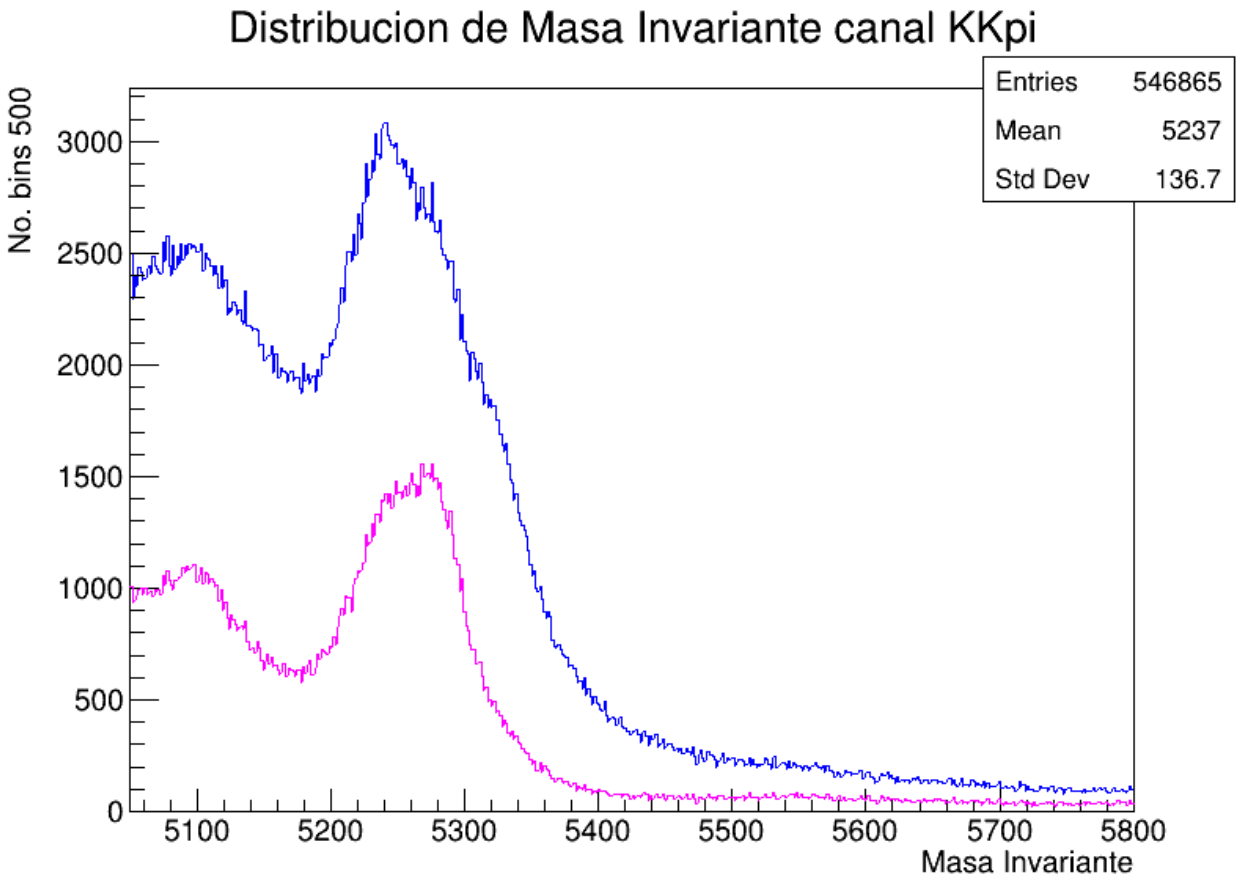


Figura 2: Distribución de $B_m$ con el corte óptimo de BDT. Magenta: Distribución de probabilidad de $B_m$ con el corte óptimo de BDT & PID del decaimiento $B^+ \to K^+K^-\pi^+$.

Cuando se aplica el criterio de selección PID para $B^+ \rightarrow K^+K^-\pi^+$, se aprecia que hay una reducción de ruido de mis-identificación y comparado con el corte de BDT, la diferencia se ve mayor, sin embargo, la distribución propia de los datos mejoro.

## Bibliografía


A. A. Alves, J. e. (2018). *The LHCb detector at the LHC.*

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* New York: Springer.

Byron P. Roe, H.-J. Y. (2008, Febrero 2). Boosted Decision Trees as an Alternative to Artificial Neural Networks for Particle Identification. *arXiv: Physics, 3.*

K. Albertsson, S. G. (2020). *TMVA 4 Toolkit for Multivariate Data Analysis with ROOT (Users Guide).* Germany: arXiv:physics.

Lista, L. (2015). *Statistical Methods for Data Analysis in Particle Physics.* Springer.