

Facultad de Ciencias

USO DE TÉCNICAS ML COMO MÉTODO ALTERNATIVO PARA LA IDENTIFICACIÓN DE JETS ADICIONALES EN EVENTOS TOP

(Use of ML techniques as an alternative method of additional jets identification on top events)

> Trabajo de Fin de Grado para acceder al

GRADO EN FÍSICA

Autor: Leopoldo Cebreiro Martinez

Director: Javier Brochero Cifuentes

Septiembre - 2023

Agradecimientos

Mi estancia en Santander ha sido muy fructífera, no solo he crecido como persona si no que he hecho grandes amistades. Al principio dudaba si moverme a una universidad o quedarme en Galicia, pero ahora puedo decir que he tomado la decisión correcta. Con respecto al ámbito académico, el comportamiento de la naturaleza siempre me ha interesado, siempre he visto el mundo y su naturaleza como algo bello y hermoso, estoy muy agradecido de formar parte de él y mi manera de mostrar ese aprecio es intentando desentrañar su misterio. Estos años estudiando física fueron duros sin duda, pero el conocimiento que he ganado y las herramientas que he aprendido han merecido la pena.

Lo primero dar las gracias a mi director del trabajo Javier Brochero, ha sido un gran director respondiéndome a todas mis dudas y guiándome en el trabajo. Ha sabido explicarme conceptos con claridad, siempre ha estado disponible en cualquier momento y ha tenido mucha paciencia conmigo. Sin duda un placer trabajar con él.

En estos años he hecho grandes amistades que voy a llevar en mi corazón por siempre. He conocido a gente que me ha comprendido a la perfección con los que he podido reír y compartir buenos momentos. Gracias a todos mis compañeros de la residencia Torres Quevedo por vuestra amistad, recordare siempre los buenos momentos dentro y fuera de la residencia, incluso compraré algún sobre Pokémon de vez en cuando a ver si hay suerte.

Gracias a todos mis compañeros de física por acogerme en su casa para disfrutar de las deliciosas paellas, tortillas y postres. Los domingos no volverán a ser lo mismo sin la reunión familiar semanal, os llevare siempre en el corazón y recordaré los momentos que pasamos juntos con dulzura.

Por último, quiero dar las gracias a la persona más importante de mi vida, gracias, Madre. Tu siempre me has apoyado y creído en mi pasara lo que pasara, siempre que he tenido un problema has hecho todo en tu poder para intentar ayudarme. No se que me deparara el futuro, pero sé que siempre estarás ahí para mi pase lo que pase, te quiero mucho.

Resumen

El objetivo de este trabajo es comprobar la eficiencia de una técnica de Machine Learning (ML) conocida como Boosted Decision Trees (BDT) para identificar jets adicionales en decaimientos de par top anti-top $(t\bar{t})$, usando simulaciones de datos del detector Compact Muon solenoid (CMS) del acelerador de partículas Large Hadron Collider (LHC) en el 2017 con energías en el centro de masas de 13 TeV. Se han empleado eventos $t\bar{t}$ con el objetivo entrenar varias BDT para que identifiquen los jets adicionales. Se ha comparado la eficiencia de la técnica de ML con el método canónico usado en la física del top: la reconstrucción cinemática. Las eficiencias de identificación de jets adicionales en las BDT han alcanzado máximos de hasta el 57 % mientras que la reconstrucción cinemática ha alcanzado máximos del 75 %.

Palabras clave: Quark Top, jets, Kinfitter, BDT

Abstract

The objective of this work is to assess the effectiveness of Machine Learning (ML) techniques known as Boosted Decision Trees (BDT) for the identification of additional jets in decays of top anti-top pairs $(t\bar{t})$, by means of data simulations from the Compact Muon Solenoid (CMS) detector at the Large Hadron Collider (LHC) in 2017, operating at a centerof-mass energy of 13 TeV. $t\bar{t}$ decay data have been employed to train multiple BDT aimed at identifying the additional jets. The efficiency of the ML technique has been compared against the canonical method used in top physics: kinematic reconstruction. The identification efficiencies of additional jets using BDT have reached maximum values of up to 57%, whereas kinematic reconstruction has achieved maximum efficiencies of 75%. **Keywords:** Top Quark, jets, Kinfitter, BDT

/		
T	11	•
In	n	ICP
	u.	IUU

Ín	dice	de figuras	4			
Ín	dice	de tablas	6			
1.	Intr	oducción	7			
2.	Mai	rco Teórico	8			
	2.1.	El Modelo Estándar	8			
	2.2.	El quark top	9			
		2.2.1. Decaimiento del quark top	9			
		2.2.2. Categorias $t\bar{t}$	10			
	2.3.	El Consejo Europeo para la Investigación Nuclear (CERN): El Gran Colisionador				
		de Hadrones (LHC) y el Compact Muon Solenoid (CMS)	11			
	2.4.	Software ROOT: Toolkit for Multivariate Data Analysis (TMVA) Package	12			
		2.4.1. Reconstrucción cinemática y Boosted Decision Trees (BDT)	12			
3.	Figu	uras de Control y entrenamiento ML	14			
	3.1.	Figuras de control y variables de entrenamiento	14			
		3.1.1. Figuras de control para jets singulares	14			
		3.1.2. Figuras de control para pares de jets	19			
	3.2.	Entrenamiento ML con jets singulares	28			
	3.3.	BDT usando pares de Jets y optimizaciones	30			
		3.3.1. BDT us ando como fondos el bosón W y el quark Top \hfill	30			
		3.3.2. BDT usando de fondo en bosón W con corte a la masa invariante	33			
		3.3.3. BDT combinando los fondos top y W \hdots	34			
		3.3.4. BDT combinando los fondos top y W con corte a la masa invariante $\ . \ .$	36			
		3.3.5. BDT combinando los fondos top y W con corte a la masa invariante y al				
		CSV	37			
	3.4.	Sobre-ajuste	39			
4.	\mathbf{Res}	ultados:	40			
	4.1.	Emparejamientos Gen-Reco	40			
	4.2.	Eficiencias				

5. Conclusión

Referencias

 $\mathbf{45}$

43

Índice de figuras

1.	Partículas del Modelo Estándar	8
2.	Diagrama de Feymann para los tipos de decaimiento $t\bar{t}$	10
3.	Esquema del CMS	12
4.	Distribución de la energía del jet para el fondo	15
5.	Distribución de la energía del jet para las señal	15
6.	Distribución de η del jet para el fondo	16
7.	Distribución de η del jet para la señal $\ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots$	16
8.	Distribución del momento transversal del jet para el fondo	17
9.	Distribución del momento transversal del jet para la señal	17
10.	Distribución de ϕ del jet para el fondo 	18
11.	Distribución de ϕ del jet para la señal $\hfill \ldots \hfill \hfill \ldots \hfill \ldots \hfill \hfill \hfill \hfill \hfill \ldots \hfill \hf$	18
12.	Distribución de CSV del jet para el fondo	19
13.	Distribución de CSV del jet para el fondo	19
14.	Distribución de $\Delta \eta$ de la pareja de jets para el fondo top	21
15.	Distribución de $\Delta \eta$ de la pareja de jets para el fondo W	21
16.	Distribución de $\Delta \eta$ de la pareja de jets para la señal $\ldots \ldots \ldots \ldots \ldots \ldots$	22
17.	Distribución de $\Delta \phi$ de la pareja de jets para el fondo top $\ldots \ldots \ldots \ldots$	22
18.	Distribución de $\Delta\phi$ de la pareja de jets para el fondo W	23
19.	Distribución de $\Delta \phi$ de la pareja jets para la señal	23
20.	Distribución de ΔR de la pareja de jets para el fondo top $\ldots \ldots \ldots \ldots$	24
21.	Distribución de ΔR de la pareja de jets para el fondo W	24
22.	Distribución de ΔR de la pareja de jets para la señal	25
23.	Distribución de la masa invariante (InvMass) de la pareja de jets para el fondo top	25
24.	Distribución de la masa invariante (Inv Mass) de la pareja de jets para el fondo W	26
25.	Distribución de la masa invariante ($\tt InvMass$) de la pareja de jets para la señal $% (\tt InvMass)$.	26
26.	Distribución de la suma del CSV (SumCSV) de la pareja de jets para el fondo top	27
27.	Distribución de la suma del CSV (${\tt SumCSV})$ de la pareja de jets para el fondo W .	27
28.	Distribución de la suma del CSV (SumCSV) de la pareja de jets para la señal	28
29.	Variables de entrada para la BDT de prueba con jets individuales	29
30.	Matrices de correlación	29
31.	BDT de prueba	30

32.	Variables de entrada usando el fondo W	31
33.	Respuesta de la BDT usando el fondo W	32
34.	Variables de entrada usando el fondo Top	32
35.	Respuesta de la BDT usando el fondo top	33
36.	Variables de entrada usando el fondo W con corte a la masa invariante	34
37.	Respuesta de la BDT usando el fondo W con corte a la masa invariante $\ .\ .\ .$	34
38.	Variables de entrada para la combinación de fondos top y W \hdots	35
39.	Respuesta de la BDT para la combinación de fondos top y W	36
40.	Variables de entrada para la combinación de fondos top y W con corte a la masa	
	invariante	36
41.	Respuesta de la BDT para la combinación de fondos top y W con corte a la masa	
	invariante	37
42.	Variables de entrada combinando los fondos top y W con corte a la masa invariante	
	y al CSV	38
43.	Respuesta de la BDT combinando los fondos Top y W con corte a la masa inva-	
	riante y al CSV	38

Índice de tablas

1.	Número de jets adicionales dependiendo del tipo de decaimiento $\ldots \ldots \ldots$	10
2.	Eficiencias obtenidas usando la BDT entrenada con fondo W	41
3.	Eficiencias obtenidas usando la BDT entrenada con fondo top 	41
4.	Eficiencias obtenidas usando la BDT entrenada con fondo W y corte a la masa	
	invariante	41
5.	Eficiencias obtenidas usando la BDT entrenada con la combinación de fondos top	
	y W	42
6.	Eficiencias obtenidas usando la BDT entrenada con la combinación de fondos top	
	y W con corte a la masa invariante	42
7.	Eficiencias obtenidas usando la BDT entrenada con combinación de fondos top y	
	W con corte a la masa invariante y a la SumCSV	42

1. Introducción

El Modelo Estándar es una teoría cuántica que postula la composición de la materia como una agrupación de constituyentes puntuales sin estructura llamados leptones y quarks. Dentro de los quarks está el más pesado de todos: el quark top. Su relevancia en la física de partículas actual es debido a que la producción de pares top anti-top $(t\bar{t})$ está presente como fondo en muchos procesos del Modelo Estándar. De igual forma, la importancia de estos eventos reside en que aparecen en multitud de ámbitos de la física de partículas: están presentes como fondo en muchos experimentos, sirven para hacer calibraciones, validar la teoría del Modelo Estándar y estudiar la misma partícula top. La vida media del quark top es de ~ 10^{-25} s por lo que decae rápidamente antes de hadronizar. En este decaimiento se generan un quark b y un bosón W por cada quark Top, el bosón W puede tener dos tipos de decaimientos donde se generan leptones o quarks. Los eventos $t\bar{t}$ suelen ir acompañados de chorros de partículas adicionales (en inglés: jets), esto se conoce como $t\bar{t}$ +jets.

El objetivo de este documento es utilizar técnicas de aprendizaje automático (en inglés: *Machine Learning*, ML) como método alternativo de identificación de jets adicionales en eventos $t\bar{t}$; la técnica de ML empleada se conoce como *Boosted Decision Trees* (BDT). Se han usado simulaciones de eventos $t\bar{t}$ con los que se han entrenado las BDT para ser capaces de identificar correctamente los jets. La eficiencia de estos métodos se comparará con la eficiencia del método canónico utilizado en física del top: la reconstrucción cinemática conocida como *Kinfitter* (de su nombre en inglés: *Kinematic Fitter*).

Para estudiar estos productos del decaimiento se utilizan detectores como el *Compact Muon Solenoid* (CMS) que identifican los productos del decaimiento como leptones y jets, entre otros. Por medio de algoritmos y con los datos recogidos por el detector, se pueden reconstruir las partículas originales y así hallar propiedades como su sección eficaz o su masa.

El primer apartado de este trabajo es un marco teórico, dividido en cuatro partes. La primera parte es una pequeña introducción al Modelo Estándar de la física y resume las partículas que lo componen. La segunda parte habla de la física del quark Top, el decaimiento $t\bar{t}$ y los jets asociados a dicho decaimiento. La tercera parte habla del Gran Colisionador de Hadrones construido por el CERN y explica de manera resumida las partes del detector CMS. La cuarta parte explica tanto el programa de aprendizaje automático que se ha empleado como método alternativo para la identificación de los jets como el método de reconstrucción cinemática.

En el segundo apartado se presentan y discuten las variables. Las variables de entrada son los datos resultantes de las simulaciones de eventos $t\bar{t}$ que contienen los parámetros con los que entrenar las BDT presentados en forma de histogramas. Hay dos sets de datos: los datos resultado de hacer entrenamientos buscando identificar jets individualmente y los datos resultado de hacer entrenamientos buscando identificar parejas de jets. Con los datos para jets individuales se ha hecho una primera prueba de entrenamiento con el objetivo de familiarizarse con el funcionamiento del programa. Posteriormente, con los datos para parejas de jets se han entrenado seis instancias de las BDT cada una con diferentes alteraciones a las variables de entrada (cortes) para intentar conseguir la mejor respuesta de la BDT posible. También se discute si ha ocurrido sobre-entrenamiento en las BDT.

El tercer apartado presenta las eficiencias resultantes a la hora de identificar jets adicionales en eventos $t\bar{t}$, se muestran las eficiencias de los seis entrenamientos de las BDT y las eficiencias resultantes de la reconstrucción cinemática. Se discuten las eficiencias para cada entrenamiento de la BDT y se comprueba si las modificaciones que se han realizado a las variables de entrada han tenido impacto en las eficiencias resultantes.

El cuarto apartado está dedicado a las conclusiones, se presenta una recapitulación de la eficiencia del aprendizaje automático como método alternativo para la identificación de jets adicionales.

2. Marco Teórico

2.1. El Modelo Estándar

Entender de que esta compuesta la materia siempre ha sido una pregunta que la humanidad ha intentado responder, ya los griegos intentaron dar sentido a la composición de la materia creyendo que toda esta estaba conformada por unidades muy pequeñas e indivisibles de ahí el termino átomo que viene del griego "*atomos*"¹. Si se adelanta un poco el cronómetro la época moderna llega y con ello nuevas teorías como los modelos de Thompson y Rutherford que explican la física de dichos átomos descubriendo los electrones y los núcleos. A medida que se producen avances científicos se empieza a observar que el átomo esta compuesto de partículas aún más pequeñas: protones, neutrones y electrones. Dichos protones y neutrones están a su vez compuestos de partículas más pequeñas de ahí que se necesite un nuevo modelo para explicar su funcionamiento, he aquí la creación del Modelo Estándar.

El Modelo Estándar (del inglés: *Standard Model*, SM) surge en 1970 [1] y asienta sus bases en 3 tipos de partículas: los leptones, quarks y bosones representados por la Figura 1. Los leptones y quarks son fermiones constituyentes de la materia y caracterizados por tener espín semi-entero mientras que los bosones tienen espín entero y son los mediadores entre las interacciones de los fermiones.



Figura 1: Partículas del Modelo Estándar por generaciones y sus principales magnitudes: masa, carga eléctrica y espín [Fuente: Wikipedia].

En este modelo estándar hay tres agrupaciones (parejas) de los quarks. Se empareja un quark con carga $\frac{2}{3}$: up, charmed, top (siglas: u, c, t) con otro con carga $-\frac{1}{3}$: down, strage, bottom (siglas: d, s, b) haciendo así las tres generaciones de la materia. Por supuesto cabe decir que cada quark tiene su antipartícula respectiva cuya carga es la misma pero con el signo cambiado.

Con respecto a los bosones hay 5 y son las partículas intercambiadoras en las fuerzas de interacción. La fuerza electromagnética esta mediada por el fotón y es responsable de la emisión

¹que no se puede cortar, indivisible

de luz, como el fotón no tiene masa el alcance de interacción es infinito. La fuerza débil, cuyo rango de interacción es muy corto (~ 10^{-3} fm), esta mediada por los bosones Z^0 y W^{\pm} . La fuerza fuerte responsable de la interacción entre quarks está mediada por el gluón. Por último está el bosón de Higgs responsable del mecanismo de Higgs.

En un principio los quarks no se veían como partículas si no como cantidades matemáticas con las que hacer cálculos, para ser aceptados como partículas hay tres áreas que han proporcionado pruebas de su existencia: la espectroscopia hadronica, scattering de leptones y producción de jets [2]. Los quarks, mediante la interacción fuerte, son los componentes de los hadrones (protones, piones, neutrones, etc), cuando se habla de procesos de hadronización se refiere a como interactúan los quarks con los gluones para crear dichos hadrones. Cuando dos patrones (quark y/o un gluón) son producidos en una colisión de partículas, la energía disponible es suficiente para crear nuevos quarks y antiquarks adicionales. Estos quarks y antiquarks adicionales se combinan rápidamente y forman pares de partículas llamadas mesones (un quark y un antiquark) o bariones (tres quarks).

2.2. El quark top

Dentro de esta gran familia de quarks el quark más masivo de todos, el quark top, es el que más importancia cobra en este trabajo. Su existencia se empezó a considerar en los años 70 con la llegada del leptón tau (τ) [3]. Este nuevo fermión indicaba la existencia de una tercera generación de quarks. El primer quark encontrado, el bottom, fue descubierto en el laboratorio DESY. Este estaba caracterizado por tener una carga de $Q = -\frac{1}{3}$ así que debería existir otro quark de carga $Q = \frac{2}{3}$ que completase la generación. Este quark seria el quark top y su existencia fue confirmada en 1995 por el Fermilab [4] siendo este el ultimo quark en ser identificado y además el más masivo siendo su masa de unos 174*GeV*.

2.2.1. Decaimiento del quark top

Debido a que es tan pesado, su tiempo de vida es extremadamente corto, por ello el quark top inmediatamente decae a un quark q que puede ser d, b o s y un bosón W pero la estadística dicta que decaimiento del top quark es casi exclusivamente (99,8%) a una pareja Wb.

$$t \to q + W$$
 $q = d, b, s$

Para un par $t\bar{t}$, el resultado del decaimiento van ser dos quarks b y dos bosones W. Lo que marca la diferencia en estos decaimientos son los bosones W, como se muestra en la Figura 2 el bosón W^+ esta asociado al quark \bar{t} y el bosón W^- esta asociado al quark t. Si ambos bosones decaen a un par $q\bar{q}$ entonces el decaimiento es hadrónico, si ambos decaen a un leptón y un neutrino entonces el decaimiento es leptónico y si uno decae hadronicamente y el otro leptonicamente entonces el decaimiento es semi-leptónico. A continuación se muestran enumerados los decaimientos que pueden suceder:

- 1. Hadrónico: Ambos bosones W se desintegran en un par de quarks $q\bar{q}$.
- 2. Leptónico: Ambos bosones W decaen en un leptón y un neutrino.
- 3. Semi-leptónico: Un bosón W decae hadronicamente y el otro leptonicamente.

Como ya se ha explicado antes, en los decaimientos $t\bar{t}$ se producen quarks o leptones dependiendo del modo de decaimiento, pero lo que se registra en el detector son jets. La diferencia principal entre un quark y un jet es que un quark es una partícula elemental que forma los componentes básicos de los hadrones, mientras que un jet es un conjunto de partículas que resultan de la fragmentación y hadronización de quarks y gluones producidos en la colisión. Los jets son observables experimentales, se detectan como chorros de partículas cuyas propiedades están relacionadas con el quark que originó ese jet. En los decaimientos $t\bar{t}$ siempre van a haber

un mínimo de 2 jets procedentes de los quarks b en un decaimiento leptónico y un máximo de 6 jets si los bosones W decaen hadronicamente; si hay un decaimiento semi-leptonico entonces serán 4 jets. En la Tabla 1 se muestra el número de jets según el decaimiento mientras que la Figura 2 representa el diagrama de Feymann del decaimiento seimi-leptónico del $t\bar{t}$.

Decaimiento	N_b	N_l	$N_{q\overline{q}}$	N_{jets}
Hadronico	2	0	2	6
Leptonico	2	2	0	2
Semi-leptónico	2	1	1	4

Tabla 1: Tabla que representa el número de jets adicionales dependiendo del tipo de decaimiento. La segunda columna (N_b) indica el número de quarks *b* por evento, la tercera (N_l) el número de pares leptón-neutrino, la cuarta $(N_{q\bar{q}})$ el número de pares $q\bar{q}$, y la última (N_{jets}) el número de jets asociados a la desintegración del par según el modo de decaimiento (suma de las columnas segunda más el doble de la cuarta, al haber dos jets por cada par $q\bar{q}$).





Figura 2: Se muestra el diagrama de Feymann del decaimiento $t\bar{t}$. De manera ordenada son: leptónico (a), semi-leptónico(b) y hadrónico (c) [5].

2.2.2. Categorias $t\bar{t}$

Los eventos $t\bar{t}$ pueden llevar asociados jets adicionales ligeros (en inglés: *light-flavour*,LF) o pesados (en inglés: *Heavy-flavour*,HF). Un jet liviano se refiere a un jet que está compuesto principalmente por partículas ligeras, como mesones piones y kaones. Los jets livianos tienden a ser producidos por quarks up, down y strange, que son quarks de menor masa. Los jets livianos también pueden incluir partículas leptónicas, como electrones y positrones. Un jet pesado está compuesto principalmente por partículas con mayor masa, los jets pesados son producidos por la fragmentación y hadronización de quarks bottom y charm, que son quarks más masivos en comparación con los quarks up, down y strange ver Figura 1.

Estos jets son el producto de la hadronización de quarks y para intentar reconstruir la particula madre se utilizan algotimos que intentan asignar el sabor inicial del quark en función

del detectado en los hadrones. Hay cuatro categorías de $t\bar{t}$ +jets que pueden aparecer en los decaimientos, la categoría depende del tipo de jet adicional. A continuación se muestran las categorías ordenados por según el porcentaje que representan del total de sucesos [6]:

- 1. $t\bar{t}bb$: Eventos con al menos 4 *b*-jets, dos adicionales y dos del los quarks *t*. Si se tienen eventos con 6 jets solo el 2 % van a ser de este tipo.
- 2. $t\bar{t}b$: Eventos con 2b-jets de los quarkty uno adicional. En las mismas condiciones de antes estos eventos son un $6\,\%$
- 3. $t\bar{t}c\bar{c}$: 2 jets del top y dos jets c adicionales. Suponen un 14 %
- 4. $t\bar{t}$ LF: 2 *b*-jets y cuatro adicionales livianos, en esta categoría también entran eventos sin jets adicionales o un jet *c* junto a uno liviano. Representan el 78 %

2.3. El Consejo Europeo para la Investigación Nuclear (CERN): El Gran Colisionador de Hadrones (LHC) y el Compact Muon Solenoid (CMS)

El Consejo Europeo para la Investigación Nuclear (CERN) es una organización dedicada a la investigación de física de partículas [7]. En sus instalaciones se ubica el Gran Colisionador de Hadrones (en inglés: *Large Hadron Collider*, LHC), el mayor acelerador de partículas creado. Uno de los descubrimientos más importantes del CERN en los recientes años ha sido la confirmación de la existencia del bosón de Higgs en 2012. El CERN también es pionero en computación de alto rendimiento pues la física de partículas requiere de resolver infinidad de algoritmos muy complejos por ejemplo, la reconstrucción cinemática.

El LHC se trata de un anillo de aceleradores de 27 kilómetros de circunferencia ubicado bajo tierra, en el cual se aceleran y colisionan protones y núcleos de átomos de plomo a altas energías (13 TeV para los protones y 2.76 TeV para el plomo). Estas partículas se hacen colisionar en puntos de donde se encuentran detectores, como el Solenoide Compacto de Muones (en inglés: *Compact Muon Solenoid*, CMS). Estos detectores registran las partículas resultantes de las colisiones, permitiendo a los científicos estudiar y analizar los datos para extraer información relevante como la masa de la partícula o su sección eficaz.

El CMS es uno de los detectores principales ubicados en el LHC. Está diseñado para medir y analizar las partículas y sus interacciones que resultan de las colisiones de alta energía. El detector CMS tiene una estructura cilíndrica y contiene varios componentes, como detectores de partículas, sistemas de seguimiento de travectorias y calorímetros, que registran y miden las propiedades de las partículas resultantes de las colisiones. El detector tiene 7 partes importantes representadas por la Figura3, cuyo uso es reconstruir las partículas, detectar los caminos seguidos y saber su partícula de origen: Tracker de Silicio: Es la parte más interna compuesto por 3 tipos de detectores siendo capaz de recrear los caminos de las partículas cargadas. Sistema de píxeles: Recrean el camino seguido por las partículas pesadas, a través de esta reconstrucción se puede saber de qué punto en el espacio provienen. Tracker de bandas de silicio: Se usa para medir momentos y reconstruir caminos seguidos por las partículas. Micro strip gas chambers: Se usa para discernir los caminos que han tomado las partículas en medio del mar de partículas detectadas. Calorímetro electromagnético: Mide la energía de los fotones y electrones que llegan a él. Calorímetro hadrónico: Usando centelleadores, mide la energía de hadrones y sus productos de decaimiento. Sistema de muones: Es el detector más externo y se usa para verificar los resultados del tracker de silicio.



Figura 3: Esquema del CMS [8]

2.4. Software ROOT: Toolkit for Multivariate Data Analysis (TMVA) Package

ROOT es una herramienta de análisis de datos open-source utilizada en física de partículas y física de altas energías. Esta herramienta tiene muchas funciones de análisis como los *Boosted Decision Trees* que se han utilizado en este trabajo pero también tiene disponibles herramientas como *Deep Neural Networks* (DNN). [9]

2.4.1. Reconstrucción cinemática y Boosted Decision Trees (BDT)

La reconstrucción cinemática, también conocida como *Kinfitter* (del inglés: Kinematic Fitter), son una serie de algoritmos usados para reconstruirla cinemática del evento para hallar las propiedades y trayectorias de las partículas involucradas en una colisión. El objetivo principal de *Kinfitter* es ajustar los parámetros de un modelo a los datos experimentales obtenidos en los detectores de partículas.

El *Kinfitter* se basa en un modelo cinemático predefinido que describe cómo las partículas se producen y decaen en una colisión. Este modelo compara los datos experimentales y se ajusta a los parámetros del modelo para minimizar las diferencias entre los valores predichos y los observados, más adelante se hablará de términos como generación, reconstrucción y simulaciones de eventos para identificar las partículas. El ajuste cinemático de *Kinfitter* es ser utilizado para reconstruir trayectorias de partículas, identificar su tipo (como electrones, muones, fotones, etc.), estimar su momento y energía, y determinar las masas y propiedades de las partículas involucradas en una colisión.

Este método tiene una alta eficiencia pero la carga computacional que requiere es muy pesada, de ahí que se busquen métodos alternativos que reduzcan la carga de computo. Aquí es donde entran los métodos de aprendizaje automático denominados en inglés como *Boosted Decision Trees* (BDT)

Las BDT son una técnica de aprendizaje automático utilizada en diversos campos, pero en este trabajo su interés reside en su aplicación para la física de partículas. Son un tipo de algoritmo que combina múltiples árboles (del inglés: *Trees*) de decisión para mejorar la precisión y el rendimiento del modelo. La idea básica detrás de las BDT es entrenar una serie de árboles de decisión en secuencia, donde cada árbol se construye con unas variables de entrada (en este caso

son sets de datos que contienen magnitudes de los quarks como el momento o las coordenadas angulares). En este proyecto se tiene datos de una señal y un fondo recogidos en dos *Trees*, uno con el fondo y otro con la señal pero ambos con las mismas variables para poder hacer la discriminación. El objetivo es conseguir la mejor separación posible entre estas dos señales, la BDT logra esto asignando mayor peso a las variables de entrada que hagan más discriminantes. Al combinar las predicciones de múltiples árboles, las BDT pueden distinguir relaciones no lineales y complejas entre las variables de entrada y la variable objetivo.

En este TFG se están utilizando BDT como método alternativo al *Kinfitter* para conseguir identificar jets adicionales en eventos $t\bar{t}$.

3. Figuras de Control y entrenamiento ML

3.1. Figuras de control y variables de entrenamiento

3.1.1. Figuras de control para jets singulares

Las variables de entrenamiento son sets de datos resultado de las simulaciones cuya información será empleada por los árboles de decisión para hacer la discriminación señal-fondo. Estos datos se presentan en forma de histogramas como figuras de control, cada figura muestra la relación entre la magnitud del jet y el número de eventos. El objetivo de este TFG es aislar jets adicionales en un evento $t\bar{t}$ +jets para poder identificarlos. Como se ha explicado antes, el producto del decaimiento del W pueden ser multitud de partículas diferentes cuya aparición esta dictada por la estadística del decaimiento. En estas simulaciones se recrean estos eventos $t\bar{t}$ como se muestran en la Figura 2 y las magnitudes de los jets producto de los eventos se recogen en sets de datos, estos sets de datos se dividen en datos de **señal** que contienen la información de los jets adicionales que se quieren identificar y, mediante técnicas de ML, separarlos del **fondo** que contiene los datos de el resto de productos de desintegración, es decir, contiene los jets del decaimiento del top. Las variables contenidas en los sets de datos son las propias magnitudes de los jets: su momento (pT), energía (e), sus dos coordenadas angulares ($\eta y \phi$) y el discriminante usado para identificar b-jets: *Combined Secondary Vertex* (CSV).

Tanto el momento como la energía se refieren a la suma de los momentos y las energías de todas las partículas que componen al jet. Puesto que la energía y el momento están íntimamente relacionados se espera un comportamiento similar en las figuras de control. Ambas magnitudes están medidas en GeV. Para las energías se puede observar como la mayoría de eventos están concentrados a bajas energías y a medida que la esta aumenta, se van reduciendo el número de eventos. Esto sucede para el set de datos de la señal representados por la Figura 5 y para el set de datos del fondo representados por la Figura 4. El jet pT tiene un comportamiento muy similar al la energía, puesto que el momento y la energía están muy relacionados. Vuelve a suceder lo mismo que en las Figuras 8 y 9 la mayoría de la información está muy concentrada al principio.

El ángulo azimutal (ϕ) es una de las componentes angulares en coordenadas cilíndricas que se utiliza para describir la dirección de movimiento de una partícula. El ángulo azimutal se mide en el plano transversal al haz de partículas y se toma con respecto a una dirección de referencia. Por lo general, esta dirección de referencia es el eje z del sistema de coordenadas cilíndricas. ϕ está expresado en radianes. Tanto la señal y el fondo del jet ϕ son muy similares, toman un valor alto al principio, se mantienen constantes y decaen rápidamente cuando llegan al final. El ángulo ϕ representa la dirección azimutal en un detector de partículas, la razón de que ϕ sea constante es la siguiente: En colisiones de alta energía los jets no tienen una preferencia en la dirección azimutal, puede haber partículas que se distribuyen uniformemente en todas las direcciones azimutales posibles dentro del jet, lo que resulta en una señal de ϕ constante.

La pseudorrapidez (η) es una medida relacionada con el ángulo entre la dirección de movimiento de una partícula y la dirección del haz en un sistema de coordenadas cilíndrico. Tiene en cuenta las limitaciones de la detección y la geometría de los detectores en los experimentos. η es adimensional. Los sets de datos para el fondo del jet η representados por la Figura 6 parecen tomar casi una distribución normal haciendo una suave curva con su máximo en el centro mientras que para los sets de datos para la señal señal representados por la Figura 7 aunque mantienen un mayor número de eventos en el centro, la distribución esta más repartida a lo largo de todo el rango de valores de η . También hay que denotar que para la Figura 6 el máximo es mucho mayor, en torno a unos 18000 mientras que para la Figura 7 alcanzan solo los 6000. Ambos tienen el mismo número de eventos pero debido a la distribución una tiene los eventos más acumulados en torno al 0 resultando un máximo mayor.

El CSV es un algoritmo que devuelve la probabilidad de que un jet venga de un b-quark, toma valores de 0 a 1 y su función es distinguir los jets livianos de los pesados, por ejemplo,

un jet LF tiende a tener un CSV de valor bajo mientras que los b-jets tienden a tomar valores altos. Para la señal de fondo del b-tag se observan dos picos, uno al principio y otro al final. Puesto que el jet b-tag toma valores entre 0 y 1 dependiendo de si son jets livianos o pesados, en el set de datos para el fondo representado por la Figura 12 se tienen una buena cantidad de jets pesados puesto que hay una gran acumulación de eventos cerca del 1. En el set de datos para la señal representado por la Figura 13 la mayoría de eventos están al principio y casi nada al final, esto significa que para el set de datos de la señal hay sobre todo por jets livianos.



Figura 4: Figura de control para el set de datos de la magnitud energía que se emplearán como variable de entrada de fondo. El eje y representa el número de eventos con una escala de 10^3 . El eje x representa la energía en GeV.



Figura 5: Figura de control para el set de datos de la magnitud energía que se emplearán como variable de entrada de señal. El eje y representa el número de eventos con una escala de 10^3 . El eje x representa la energía en GeV.



Figura 6: Figura de control para el set de datos de la magnitud η que se emplearán como variable de entrada de fondo. El eje y representa el número de eventos y el eje x representa los valores de η .



Figura 7: Figura de control para el set de datos de la magnitud η que se emplearán como variable de entrada de señal. El eje y representa el número de eventos y el eje x representa los valores de η .



Figura 8: Figura de control para el set de datos de la magnitud momento transversal que se emplearán como variable de entrada de fondo. El eje y representa el número de eventos con una escala de 10^3 . El eje x representa el momento transversal en GeV.



Figura 9: Figura de control para el set de datos de la magnitud momento transversal que se emplearán como variable de entrada de señal. El eje y representa el número de eventos con una escala de 10^3 . El eje x representa el momento transversal en GeV.



Figura 10: Figura de control para el set de datos de la magnitud ϕ que se emplearán como variable de entrada de fondo. El eje y representa el número de eventos y el eje x representa los valores de ϕ en radianes.



Figura 11: Figura de control para el set de datos de la magnitud ϕ que se emplearán como variable de entrada de señal. El eje y representa el número de eventos y el eje x representa los valores de ϕ en radianes.



Figura 12: Variable de entrada para el fondo b-tag. El eje y representa el número de eventos con una escala de 10^3 . El eje x representa el CSV: la probabilidad de que los jets sean pesados o livianos (0=livianos, 1=pesados).



Figura 13: Variable de entrada para la señal b-tag. El eje y representa el número de eventos con una escala de 10^3 . El eje x representa el CSV: la probabilidad de que los jets sean pesados o livianos (0=livianos, 1=pesados).

3.1.2. Figuras de control para pares de jets

Con estos sets de datos se entrenarán las BDT, esta les asignará un peso y se creará un discriminador entre señal y fondo. Usando datos para la señal y datos para los fondos, la BDT construye una respuesta que discrimine la señal y el fondo lo mejor posible.

Las distribuciones mostradas previamente son para jets individuales, anteriores estudios han comprobado que es mas eficaz hacerlo con pares de jets. Con pares de jets las variables cambian, como se están usando parejas ahora las variables de entrada son diferencias entre las magnitudes de ambos jets. Estas nuevas variables son: el plano transverso (ΔR), $\eta \neq \phi$ ($\Delta \eta$, $\Delta \phi$), la suma de los CSV (SumCSV) y la masa invariante (InvMass). Además se van a tener dos fondos, el fondo con los quarks *b* producidos por el quark top y el fondo producido por los quarks *b* de la desintegración del *W*. La variable SumCSV es la suma de los CSV de cada jet, como cada jet puede tomar valores entre 0 y 1 y hay 2 jets SumCSV, al ser la suma, se extiende desde 0 a 2.

El comportamiento del fondo en la distribución de $\Delta \eta$ para W Figura 15 y top Figuras 14 son muy similares y también mantiene esa similitud con su respectiva señal representada por

la Figura 16. Para la Figura 14 los eventos alcanzan el mínimo a los 4 mientras que para la Figura 15 alcanzan el mínimo en 2.5, para la Figura 16 el mínimo se alcanza entorno a los 5.

Los sets de datos para las variables de entrada de fondo top y W para $\Delta \phi$, representadas por las Figuras 17 y 18 respectivamente, tienen comportamientos casi opuestos. Para el top la gran parte de los eventos están concentrados en los extremos, mientras que para el W los eventos están concentrados en el rango de -1 a 1 radianes. En el set de datos de la señal representado por la Figura 19, el comportamiento es uniforme con dos pequeños picos en el medio, casi se asemeja a una combinación de los dos set de datos.

Para ΔR el comportamiento de ambas señales de fondo, es ligeramente similar. Para el el set de datos del fondo del top Figura 20 sigue una forma dentada con los eventos comprendidos entre los valores 2-3.5 radianes mientras que el comportamiento del W Figura 21 es más suave y los eventos están concentrados a menores valores, en el rango de 0.5-1.5 radianes. Para la variable de entrada de señal Figura 22, su forma se asemeja a una combinación de los dos fondos con los eventos distribuidos en un rango más extenso.

Una de las variables que probablemente tenga más peso sea la masa invariante. Si observamos los fondos, para el fondo W, Figura 24, se ve una concentración de eventos en el rango de los 100 GeV, esto es el bosón W que tiene una masa cercana a ese valor (de esto se hablará más adelante). Para el fondo del top, Figura 23, puesto que los quarks b generados por no mantienen una relación tan estrecha como los del W la curva es más suave. Los quarks b que se observan en la Figura 23 son resultado del decaimiento de un quark t y un \bar{t} , se generan en el mismo instante pero de diferentes quarks. Para el caso de los quarks los b, son originados del decaimiento hadrónico del W por eso su masa invariante está tan bien definida, esto aparece representado en los diagramas de Feymann de la Figura 2.

Por último la variable SumCSV, para el fondo W, representado por la Figura 27, el máximo de señal se encuentra hacia la izquierda mientras que en el fondo top Figura 26 el máximo se encuentra al final. Para la Figura 28 el valor del CSV está acumulado al principio pero como los entrenamientos se quieren enfocar en los b-jets en esta variable se seleccionarán solo dichos jets por lo que en las variables de entrada de la siguiente sección se verá como la Figura 28 pica en 1 en vez de en 0 ya que al seleccionar los b-jets estos jets livianos del principio se eliminan. También hay que tener en cuenta que haciendo este corte estamos eliminando gran parte de los eventos, en la Figura 28 los eventos cercanos al cero tienen su máximo en 34000 pero los jets con valor 1 apenas llegan a los 5000 eventos. Este corte para seleccionar b-jets se hará tanto para las variables de fondo como para las variables de señal.



Figura 14: Figura de control para el set de datos de la magnitud $\Delta \eta$ que se emplearán como variable de entrada de fondo top. El eje y representa el número de eventos y el eje x representa los valores de $\Delta \eta$.



Figura 15: Figura de control para el set de datos de la magnitud $\Delta \eta$ que se emplearán como variable de entrada de fondo W. El eje y representa el número de eventos y el eje x representa los valores de $\Delta \eta$.



Figura 16: Figura de control para el set de datos de la magnitud $\Delta \eta$ que se emplearán como variable de entrada de señal. El eje y representa el número de eventos y el eje x representa los valores de $\Delta \eta$.



Figura 17: Figura de control para el set de datos de la magnitud $\Delta \phi$ que se emplearán como variable de entrada de fondo top. El eje y representa el número de eventos y el eje x representa los valores de $\Delta \phi$ en radianes.



Figura 18: Figura de control para el set de datos de la magnitud $\Delta \phi$ que se emplearán como variable de entrada de fondo W. El eje y representa el número de eventos y el eje x representa los valores de $\Delta \phi$ en radianes.



Figura 19: Figura de control para el set de datos de la magnitud $\Delta \phi$ que se emplearán como variable de entrada de señal. El eje y representa el número de eventos y el eje x representa los valores de $\Delta \phi$ en radianes.



Figura 20: Figura de control para el set de datos de la magnitud ΔR que se emplearán como variable de entrada de fondo top. El eje y representa el número de eventos y el eje x representa los valores de ΔR en radianes.



Figura 21: Figura de control para el set de datos de la magnitud ΔR que se emplearán como variable de entrada de fondo W. El eje y representa el número de eventos y el eje x representa los valores de ΔR en radianes.



Figura 22: Figura de control para el set de datos de la magnitud ΔR que se emplearán como variable de entrada de señal. El eje y representa el número de eventos y el eje x representa los valores de ΔR en radianes.



Figura 23: Figura de control para el set de datos de la magnitud masa invariante que se emplearán como variable de entrada de fondo top. El eje y representa el número de eventos con escala de 10^3 y el eje x representa la masa invariante entre los dos jets en GeV.



Figura 24: Figura de control para el set de datos de la magnitud masa invariante que se emplearán como variable de entrada de fondo W. El eje y representa el número de eventos con escala de 10^3 y el eje x representa la masa invariante entre los dos jets en GeV.



Figura 25: Figura de control para el set de datos de la magnitud masa invariante que se emplearán como variable de entrada de señal. El eje y representa el número de eventos con escala de 10^3 y el eje x representa la masa invariante entre los dos jets en GeV



Figura 26: Figura de control para el set de datos del SumCSV que se emplearán como variable de entrada de fondo top. El eje y representa el número de eventos con escala de 10^3 y el eje x representa la suma de los CSV de cada jet, como cada jet puede tomar valores entre 0 y 1 y hay 2 jets SumCSV, al ser la suma, se extiende desde 0 a 2.



Figura 27: Figura de control para el set de datos del SumCSV que se emplearán como variable de entrada de fondo W. El eje y representa el número de eventos con escala de 10^3 y el eje x representa la suma de los CSV de cada jet, como cada jet puede tomar valores entre 0 y 1 y hay 2 jets SumCSV, al ser la suma, se extiende desde 0 a 2.



Figura 28: Figura de control para el set de datos del SumCSV que se emplearán como variable de entrada de señal. El eje y representa el número de eventos con escala de 10^3 y el eje x representa la suma de los CSV de cada jet, como cada jet puede tomar valores entre 0 y 1 y hay 2 jets SumCSV, al ser la suma, se extiende desde 0 a 2.

3.2. Entrenamiento ML con jets singulares

Las BDT utilizan las variables de entrada organizadas en forma de árboles e intentan maximizar la separación de señal y fondo. Para este caso las variables de entrada son las generadas en las simulaciones, estas variables son las propias magnitudes de los jets: su momento, energía, sus dos coordenadas angulares ($\eta \neq \phi$) y el CSV(jet b-tag).

Para el entrenamiento de prueba, se muestran en la Figura 29 las variables que la BDT va a emplear. La señal se representa en azul y el fondo en rojo. El objetivo es poder discriminar la señal azul del fondo rojo lo mejor posible. El peso para la discriminación en las variables se puede observar al ver como se correlacionan el fondo y la señal. Si están solapados y tienen casi el mismo comportamiento a la BDT le va a costar sepáralos (véase jet ϕ) o si en un lado el fondo mucho mayor en comparación con la señal, la BDT interpretara que ahí la mayor contribución sera de fondo por lo que le dará mas peso a esa zona a la hora de separar las señales.

La Figura 30 representa las correlaciones entre las variables, cuanto más verdes y claras menos correlaciones. Esto va a afectar a como la BDT va a emplear las variables y a cuales les va a dar más peso a la hora de hacer el corte por ejemplo, se puede observar como la variable pT y E están muy correlaciones a juzgar por el color anaranjado en la matriz, esto es de esperar pues el momento y energía están íntimamente relacionados. Esto implica que a la hora computar estas variables van a aportar casi lo mismo, así que lo eficiente seria eliminar una si se quiere optimizar el proceso.



Figura 29: Variables de entrada para la BDT de prueba que se ha hecho con jets individuales. De izquierda a derecha las variables son: momento, dos coordenadas angulares ($\eta \neq \phi$), energía y b-tag. Las gráficas se dividen en señal (azul) y fondo (rojo).



Figura 30: Matrices de correlación entre las variables momento, η , ϕ , energía y b-tag. Un color verde claro representa poca correlación entre las variables mientras que un color anaranjado representa una fuerte correlación. Las diagonales están en rojo pues la correlación es de 100 ya que son las mismas variables.

En la Figura 31 se muestra el resultado de la BDT, es un resultado poco discriminante, para separar la señal del fondo lo más recomendable es hacer el corte alrededor del 0.1 pues así se elimina el fondo. El problema es la superposición de señal y fondo, es cierto que a partir del 0.1 hay mucha más señal que fondo y si se hace el corte ahí se elimina el fondo pero también se elimina mucha información, apenas queda señal. La clave es seleccionar las variables y hacer las optimizaciones que se consideren para poder conseguir una respuesta de la BDT que claramente separe el fondo de la señal.



Figura 31: Respuesta de la BDT para jets individuales que discrimina la señal (azul) del fondo(rojo) usando las variables de la Figura 29 para los sets de datos de entrenamiento (valores puntuales) y datos de verificación (histogramas)

El trabajo en el que se apoya este trabajo [6] ya se tiene esta consideración y se observa como es más eficiente hacerlo con pares de jets en vez de con jets singulares. El resto de las pruebas se harán con pares de jets así que las variables que entran en juego ahora van a ser diferencias (deltas): el plano transverso (ΔR), $\eta y \phi$ ($\Delta \eta$, $\Delta \phi$), la suma de los CSV (SumCSV) y la masa invariante (InvMass).

3.3. BDT usando pares de Jets y optimizaciones

El hecho de hacer de nuevo las BDT usando pares de jets ya de por si es una optimización pues previos estudios [6] vieron que se mejoraba la discriminación. Ahora que se emplean pares de jets se van a tener dos fondos: el producido por el bosón W y el fondo del top. Estos fondos se refieren a los quarks b producidos en la desintegración del top.

Además, hay que optimizar para los $t\bar{t}b\bar{b}$ y el $t\bar{t}bj$ así que la condición inicial que van a llevar todas las BDT ahora es la selección de $t\bar{t}bj$ y $t\bar{t}b\bar{b}$. En total se han hecho seis instancias de BDT cada una con diferentes cortes. Se presentan organizadas en orden cronológico.

3.3.1. BDT usando como fondos el bosón W y el quark Top

Lo primero que se puede observar son las variables de las Figuras 32 y 34. En la masa invariante de la Figura 32 se puede ver un pico muy marcado al principio esto es el bosón W. Dicho bosón tiene una masa muy bien definida de unos 80 GeV [10], de ahí ese pico tan fuerte que se observa, esta masa tan bien definida se puede cortar del fondo para intentar mejorar la BDT. Hay que recalcar que la BDT no entiende de bosones y quarks solo de señal y fondo; al obviar datos ya conocidos, como la masa del W, se está optimizando la cantidad de información que se le da y por tanto se podría mejorar la separación de fondo y señal. La variable de SumCSV crece a valores bajos y está muy bien separada de la señal por lo que la BDT le va a asignar prioridad a esta variable a la hora de hacer la discriminación señal-fondo. Las otras tres variables restantes $\Delta \eta$, $\Delta \phi$ y ΔR tienen la señal y el fondo bastante superpuestos en comparación con los fondos y señales tan bien distinguidos como SumCSV y la InvMass, si bien es cierto que aportan información, la BDT le va a dar mucho más peso a las variables SumCSV y InvMass.

Para las variables del fondo top representado por la Figura 34 al contrario que con el W la variable InvMass decae de manera más suave haciendo una curva que se superpone con la señal. Esto sucede porque la masa de los quarks b que surgen del decaimiento del $t\bar{t}$ esta bastante separada. Cada quark b se genera de un quark Top independiente del otro pero en el mismo momento al contrario que los b surgidos del bosón W, que tienen relación con su decaimiento, Figura 2. De la misma manera que en el W, en la variable SumCSV su fondo y señal están muy bien separados lo que indica que esta va a ser una variable de mucho peso la diferencia es que con el top el fondo se hace mayor hacia el final. Las variables $\Delta \eta$, $\Delta \phi$ y ΔR vuelven a tomar un papel secundario aquí también.

Las BDT resultantes, representadas por las Figuras 35 para entrenamiento con fondo top y 33 para entrenamiento con fondo W, han tenido una mejora en la discriminancia señalfondo considerable, la separación está mejor definida si se compara con la Figura 31 de BDT de jets singulares en la que hay una mayor superposición señal-fondo. Esto comprueba la idea de utilizar pares de jets para obtener una mayor discriminación es acertada y eficiente.

Ahora que se disponen de dos BDT discriminantes, entonces, si se combinan las señales de fondo del W y el top se podría tener una mejora en el resultado. El siguiente paso es volver a hacer las BDT combinando la señal de fondo de top y W y aplicar el corte al bosón W.



Figura 32: Variables de entrada para la primera BDT que se ha hecho con pares de jets y usando de fondo la señal W. De izquierda a derecha las variables son: $\Delta \eta$, $\Delta \phi$, ΔR , InvMass y SumCSV. Las gráficas se dividen en señal (azul) y fondo (rojo).



Figura 33: Respuesta de la primera BDT que discrimina la señal (azul) del fondo(rojo) para pares de jets usando fondo de la señal del W y las variables de la Figura[32]



Figura 34: Variables de entrada para la segunda BDT que se ha hecho con pares de jets y usando de fondo la señal top. De izquierda a derecha las variables son: $\Delta \eta$, $\Delta \phi$, ΔR , InvMass y SumCSV. Las gráficas se dividen en señal (azul) y fondo (rojo).



Figura 35: Respuesta de la segunda BDT que discrimina la señal (azul) del fondo(rojo) para pares de jets usando fondo la señal del top y las variables de la Figura[34].

3.3.2. BDT usando de fondo en bosón W con corte a la masa invariante

En esta instancia se ha empleado como fondo la señal del W y a la variable InvMass se le ha hecho un corte de 60 a 100 GeV para eliminar la masa del bosón W (no confundir la señal del fondo con el corte que le estamos haciendo al la variable InvMass, ambos se refieren al bosón W pero son cosas distintas). Por lo que se observa en las variables Figura 36, el corte del bosón W ha bajado el pico de intensidad en la variable InvMass pero la respuesta de la BDT Figura 37 no se ha visto muy afectada.



Figura 36: Variables de entrada para la tercera BDT que se ha hecho con pares de jets y usando de fondo la señal W. De izquierda a derecha las variables son: $\Delta \eta$, $\Delta \phi$, ΔR , InvMass y SumCSV. A la variable InvMass se le ha hecho un corte entre 60 y 100 para eliminar el bosón W. Las gráficas se dividen en señal (azul) y fondo (rojo).



Figura 37: Respuesta de la tercera BDT que discrimina la señal (azul) del fondo(rojo) para pares de jets usando fondo la señal del W con el corte de 60 a 100 en la masa invariante Figura 36.

3.3.3. BDT combinando los fondos top y W

Puesto que las respuestas de la BDT con fondos de top y W parecen ser buenos, se va a hacer la combinación de ambos fondos para intentar mejorar la discriminancia señal-fondo. En la Figura 38 se representa como el juntar los fondos afecta a las variables, especialmente a las variables InvMass y SumCSV, las más discriminantes. En la variable InvMass se sigue teniendo ese pico tan marcado de la Figura 32 pero ahora esta combinado con el fondo del top, Figura 34. En la variable SumCSV sigue estando presente ese pico intenso de la señal pero ahora el fondo está concentrado tanto al principio como al final, como es de esperar al combinar las señales. Con respecto a la respuesta de la BDT Figura 39 hay una buena discriminación de señal y fondo, se tienen dos picos bastante intensos bien separados.



Figura 38: Variables de entrada para la cuarta BDT que se ha hecho con pares de jets y usando la combinación de fondos las señales top y W. De izquierda a derecha las variables son: $\Delta \eta$, $\Delta \phi$, ΔR , InvMass y SumCSV. Las gráficas se dividen en señal (azul) y fondo (rojo).



Figura 39: Respuesta de la cuarta BDT que discrimina la señal (azul) del fondo(rojo) para pares de jets usando la combinación de fondos las señales top y W para las variables de la Figura 38

3.3.4. BDT combinando los fondos top y W con corte a la masa invariante

A la combinación de fondos también se le ha aplicado el corte a la masa invariante para eliminar el bosón W. La respuesta parece empeorar un poco, si se hace un corte en el 0.1 nos estamos llevando más fondo que en la anterior, aún así se sigue teniendo una separación señal-fondo aceptable.



Figura 40: Variables de entrada para la quinta BDT que se ha hecho con pares de jets y usando la combinación de fondos las señales top y W. De izquierda a derecha las variables son: $\Delta \eta$, $\Delta \phi$), ΔR , InvMass y SumCSV. A la variable InvMass se le ha hecho un corte de 60 a 100 para eliminar el bosón W. Las gráficas se dividen en señal (azul) y fondo (rojo).



Figura 41: Respuesta de la quinta BDT que discrimina la señal (azul) del fondo(rojo) para pares de jets usando la combinación de fondos las señales top y W para las variables de la Figura 40. Se le ha hecho un corte a la masa invariante de 60 a 100 para eliminar el bosón W

3.3.5. BDT combinando los fondos top y W con corte a la masa invariante y al CSV

Esta BDT tiene la combinación de fondos, el corte en la masa invariante y además se le ha fijado la variable SumCSV para que coja solo valores mayores o iguales a uno de este modo se fuerza a la BDT a evaluar a los b-jets y olvidarse de los jets livianos. El resultado de esta BDT representado por la Figura 43 es muy satisfactorio, no solo presenta dos picos bien definidos si no que la separación entre fondo y señal está muy bien distinguida, cuando la señal llega al 0 ya es casi mínima y decae rápidamente hasta que en el 0.2 apenas es notable.



Figura 42: Variables de entrada para la sexta BDT que se ha hecho con pares de jets y usando la combinación de fondos las señales top y W. De izquierda a derecha las variables son: $\Delta \eta$, $\Delta \phi$, ΔR , InvMass y SumCSV. A la variable InvMass se le ha hecho un corte de 60 a 100 para eliminar el bosón W y a la variable SumCSV se ha modificado para que solo seleccione valores de fondo mayores o iguales a 1. Las gráficas se dividen en señal (azul) y fondo (rojo).



Figura 43: Respuesta de la sexta BDT que discrimina la señal (azul) del fondo(rojo) para pares de jets usando la combinación de fondos las señales top y W para las variables de la Figura 42. A la variable InvMass se le ha hecho un corte de 60 a 100 para eliminar el bosón W y a la variable SumCSV se ha modificado para que solo seleccione valores de fondo mayores o iguales a 1.

3.4. Sobre-ajuste

Antes de terminar este capítulo es de interés discutir que en ningún caso se ha producido un sobre-ajuste (*overtraining*) en el aprendizaje automático. El sobre-ajuste sucede cuando un modelo se ajusta demasiado bien a los datos de entrenamiento y pierde su capacidad de discernir nuevos datos. En los entrenamientos se observaría *overtraining* si las respuestas de la BDT no se ajustasen a las curvas de puntos con errores, es decir, se habría producido un sobre-ajuste si los sets de datos de entrenamiento no concordasen con los datos de verificación, los histogramas.

Para comprobar las optimizaciones, en el siguiente apartado, se hallarán las eficiencias del *Kinfitter* y las BDT. Se discutirá que entrenamiento ha dado mejores eficiencias y si la BDT es una alternativa competitiva a la reconstrucción cinemática.

4. Resultados:

4.1. Emparejamientos Gen-Reco

Se están simulando eventos que producen señal $t\bar{t}$ +jets. Estas simulaciones pueden ser de diferentes niveles a nivel Generación (Gen) o Reconstrucción (Reco): GenParticles, GenJets o RecoJets [6]:

- 1. GenParticles: Simulación teniendo en cuenta la física que hay detrás del problema. Cada evento de desintegración tiene asociados 6 jets a los cuales se les asigna un valor (ID) del 1 al 6 indicando que partículas es y sus principales magnitudes (energía, momento, etc.). En orden son. d: 1, u: 2, s: 3, c: 4, b: 5, t: 6
- 2. GenJets: En la simulación de este caso se desarrolla la hadronización del quark, la partícula conserva la identificación del quark que la produjo y conservan sus magnitudes.
- 3. RecoJets: En esta simulación se tienen en cuenta la interacción de las partículas y jets con el detector.

Las tablas que se muestran en la siguiente sección son los resultados (eficiencias) de las BDT que se han entrenado y la del *Kinfitter* a la hora de poder identificar los jets adicionales para cada categoría. En todas las tablas la primera fila representa la eficiencia del emparejamiento entre GenJets y RecoData, que en ningún caso es del 100 % . Para dar un ejemplo práctico: Si se tiene una eficiencia del emparejamiento entre GenJets y RecoData para $t\bar{t}b\bar{b}$ del 80 % entonces de 100 eventos 80 han sido emparejados correctamente y el 20 restante no tuvo emparejamiento ya sea por una mala reconstrucción (que no se consigan emparejar los 4 quark del evento) o porque no han sido detectados por el detector. Esto significa que para el resto de los resultados estamos trabajando con esa eficiencia, siguiendo el ejemplo anterior: si tenemos una eficiencia en la BDT del 50 % va a ser con respecto a esos 80 eventos, es decir, 40 eventos han emparejado y los otros 40 no.

4.2. Eficiencias

Una vez obtenidas todas las BDT se correrán en el programa que analiza los GenJets y los RecoJets, los emparejará y dará como resultado las eficiencias *Gen-Reco* de las BDT, de las BDT para reconocer al menos un jet y además de la eficiencia del *Kinfitter* en cada categoría para poder comparar los dos métodos.

Las mejores eficiencias surgen de combinar los fondos del top y del W, Tabla 5 y Tabla 6 respectivamente, se han obtenido eficiencias de entorno al 50 % en ambos casos. Utilizando los fondos por separado, Tabla 2 para el fondo W y Tabla 3 para el fondo top, se obtienen eficiencias de alrededor del 40 %. Como se ha discutido en el análisis, el combinar los fondos ha mejorado la eficiencia esto se ha comprobado con los resultados obtenidos. Es de interés discutir que quitar de la variable InvMass el bosón W no ha afectado mucho al resultado, las eficiencias de la Tabla 5 y Tabla 6 para cada categoría apenas se ha visto afectada. Teniendo en cuenta la Figura 38 el motivo de que las eficiencias no se vean demasiado alteradas es porque al combinar señales del top y el W, el fondo queda más distribuido y quitarle el bosón W en un rango tan escueto afecta poco si se considera el fondo total.

En la Tabla 4 la eficiencia decae en comparación con la Tabla 2 en esta situación puede ser que el hecho de que en la BDT para el fondo W, Figura 32, la variable InvMass está muy concentrada en el rango de 60 a 100 por lo que la BDT le va a asignar mucha prioridad y al quitarle el bosón W, justo en ese rango, la variable InvMass, Figura 36, pierde poder de discriminación y por tanto baja la eficiencia.

Las eficiencias para la Tabla 7 decaen, el corte en el CSV no es muy adecuado. Este corte se hizo para intentar eliminar los jets livianos y enfocarse en los b-jets pero en el resultado final se ve que tomar este corte no es lo mejor.

La ultima fila de valores de cada tabla se refiere a la eficiencia al detectar al menos un jet adicional, en general son altas, la condición de detectar al menos un jet es difícil de cumplir. Habiendo detectado un jet, es posible hacer una segunda iteración a partir de la configuración de la primera enfocando la BDT al segundo jet.

Para terminar esta sección, se ha visto que el método *Kinfitter* sigue siendo el método con mejores resultados pero la BDT optimizada aunque con peor eficiencia puede ser una alternativa. Si bien es cierto que las discrepancias en las eficiencias son del 20 %, la BDT es mucho más rápida así que puede ser una opción cuando no se dispone de mucho poder computacional o no se dispone de mucho tiempo. La mejor optimización ha sido la resultante de combinar las BDT para cada fondo, las eficiencias de las BDT con fondos individuales son de al rededor del 40 % . Se ha comprobado que combinar los fondos ha dado una mejora de entre el 10 % y el 20 % dependiendo del canal estudiado. A continuación se muestran las tablas:

Eficiencias	ttbb	ttbj	ttcc	ttLF
emparejamiento	0.82	0.85	0.90	0.90
kinfitter	0.69	0.71	0.74	0.75
BDT	0.48	0.44	0.35	0.32
BDT al menos 1 jet	0.68	0.68	0.66	0.63

Tabla 2: Eficiencias obtenidas obtenidas usando la BDT entrenada con fondo W [3.3.1] para cada canal, $t\bar{t}b\bar{b}$, $t\bar{t}c\bar{c}$ y $t\bar{t}LF$. Por orden de filas las eficiencias son: Primera fila: eficiencia del emparejamiento GenJets-RecoData. Segunda fila: eficiencia de la reconstrucción cinemática (*Kinfitter*). Tercera fila: eficiencia de la BDT. Cuarta fila: eficiencia de la BDT en reconocer al menos 1 jet.

Eficiencias	ttbb	ttbj	ttcc	ttLF
emparejamiento	0.82	0.86	0.90	0.90
kinfitter	0.69	0.71	0.74	0.75
BDT	0.40	0.41	0.57	0.57
BDT al menos 1 jet	0.49	0.59	0.55	0.58

Tabla 3: Eficiencias obtenidas obtenidas usando la BDT entrenada con fondo top [3.3.1] para cada canal, $t\bar{t}b\bar{b}$, $t\bar{t}c\bar{c}$ y $t\bar{t}LF$. Por orden de filas las eficiencias son: Primera fila: eficiencia del emparejamiento GenJets-RecoData. Segunda fila: eficiencia de la reconstrucción cinemática (*Kinfitter*). Tercera fila: eficiencia de la BDT. Cuarta fila: eficiencia de la BDT en reconocer al menos 1 jet.

Eficiencias	ttbb	ttbj	ttcc	ttLF
emparejamiento	0.82	0.86	0.90	0.90
kinfitter	0.69	0.72	0.74	0.75
BDT	0.45	0.38	0.28	0.27
BDT al menos 1 jet	0.59	0.57	0.45	0.50

Tabla 4: Eficiencias obtenidas obtenidas usando la BDT entrenada con fondo W y corte a la masa invariante entre 60 y 100 GeV [3.3.2] para cada canal, $t\bar{t}b\bar{b}$, $t\bar{t}bj$, $t\bar{t}c\bar{c}$ y $t\bar{t}LF$. Por orden de filas las eficiencias son: Primera fila: eficiencia del emparejamiento GenJets-RecoData. Segunda fila: eficiencia de la reconstrucción cinemática (*Kinfitter*). Tercera fila: eficiencia de la BDT. Cuarta fila: eficiencia de la BDT en reconocer al menos 1 jet.

Eficiencias	ttbb	ttbj	ttcc	ttLF
emparejamiento	0.82	0.86	0.89	0.90
kinfitter	0.69	0.72	0.74	0.74
BDT	0.51	0.51	0.52	0.53
BDT al menos 1 jet	0.63	0.63	0.63	0.67

Tabla 5: Eficiencias obtenidas obtenidas usando la BDT entrenada con la combinación de fondos top y W [3.3.3] para cada canal, $t\bar{t}b\bar{b}$, $t\bar{t}c\bar{c}$ y $t\bar{t}LF$. Por orden de filas las eficiencias son: Primera fila: eficiencia del emparejamiento GenJets-RecoData. Segunda fila: eficiencia de la reconstrucción cinemática (*Kinfitter*). Tercera fila: eficiencia de la BDT. Cuarta fila: eficiencia de la BDT en reconocer al menos 1 jet.

Eficiencias	ttbb	ttbj	ttcc	ttLF
emparejamiento	0.82	0.86	0.89	0.90
kinfitter	0.69	0.72	0.74	0.75
BDT	0.51	0.52	0.53	0.51
BDT al menos 1 jet	0.62	0.62	0.64	0.68

Tabla 6: Eficiencias obtenidas obtenidas usando la BDT entrenada con combinación de fondos top y W con corte a la masa invariante entre 60 y 100 GeV [3.3.4] para cada canal, $t\bar{t}b\bar{b}$, $t\bar{t}bj$, $t\bar{t}c\bar{c}$ y $t\bar{t}LF$. Por orden de filas las eficiencias son: Primera fila: eficiencia del emparejamiento GenJets-RecoData. Segunda fila: eficiencia de la reconstrucción cinemática (*Kinfitter*). Tercera fila: eficiencia de la BDT. Cuarta fila: eficiencia de la BDT en reconocer al menos 1 jet.

Eficiencias	ttbb	ttbj	ttcc	ttLF
emparejamiento	0.82	0.86	0.89	0.90
kinfitter	0.69	0.72	0.74	0.74
BDT	0.27	0.27	0.29	0.29
BDT al menos 1 jet	0.34	0.35	0.37	0.39

Tabla 7: Eficiencias obtenidas obtenidas usando la BDT entrenada con combinación de fondos top y W con corte a la masa invariante entre 60 y 100 GeV y a la SumCSV para que tome solo los valores mayores o iguales a 1 [3.3.5] para cada canal, $t\bar{t}b\bar{b}$, $t\bar{t}bj$, $t\bar{t}c\bar{c}$ y $t\bar{t}LF$. Por orden de filas las eficiencias son: Primera fila: eficiencia del emparejamiento GenJets-RecoData. Segunda fila: eficiencia de la reconstrucción cinemática (*Kinfitter*). Tercera fila: eficiencia de la BDT. Cuarta fila: eficiencia de la BDT en reconocer al menos 1 jet.

Para terminar esta sección se hablará sobre el tiempo de cómputo. El *Kinfitter* es un algoritmo que utiliza las propiedades físicas de la partícula para hacer una reconstrucción, es el método que más se utiliza ya que es el más eficientes da, como se ha comprobado. El problema de este algoritmo es el tiempo y esfuerzo computacional que requiere. Para intentar aliviar esto se han empleado las BDT que requieren un menor esfuerzo computacional y tiempo. El total de tiempo de cómputo (para calcular las eficiencias del *Kinfitter* y de las BDT) para cada categoría ha sido de: 30 minutos para $t\bar{t}b\bar{b}$, para $t\bar{t}bj$ 50 minutos y para $t\bar{t}c\bar{c}$ y $t\bar{t}LF$ alrededor de 2 horas. En este trabajo se han usado las CPUs que ha dispuesto la universidad, si se utilizasen GPUs el tiempo habría disminuido al ser estas más potentes.

5. Conclusión

Se han realizado los estudios necesarios para seleccionar eventos $t\bar{t}$ en el canal de decaimiento l+jets usando simulaciones de ML de colisiones protón-protón a una energía de centro de masa de 13 TeV. El ML se pasa a través de la simulación del detector CMS para mayor veracidad. En este tipo de eventos estudiamos un método alternativo para la identificación de jets adicionales a aquellos provenientes del decaimiento del par $t\bar{t}$. La técnica empleada se denomina *Boosted Decision Tree*.

Como primera prueba, se ha hecho un entrenamiento de con jets individuales para observar como la BDT es capaz de discernir fondos y señales a partir del peso discriminante que les asigna a las variables de entrada. En las correlaciones se observó que puesto que la energía y momento mantenían alta relación, como es lógico, una se puede eliminar para evitar dar dos veces la misma información y mejorar el resultado.

Una vez hechas las primeras pruebas se paso al entrenamiento de BDT usando pares de jets para mejorar el resultado. Primero se examinó como respondían las BDT al darle como fondo la señal del top y la señal del W. La respuesta usando como fondo el W se intento mejorar quitándole el fondo del bosón W a la variable de masa invariante. Puesto que las BDT top y Wse obtuvieron resultados con buena discriminancia se paso a la combinación de ambos fondos para intentar mejorar el resultado aún más. Una vez hecha esta combinación se procedieron a hacer cortes en las variables de entrada para otra vez mejorar el resultado. Se le hizo el corte del bosón W y luego se volvió a iterar haciendo una nueva BDT con un corte a mayores en la variable SumCSV para intentar eliminar jet livianos y forzar el entrenamiento en b-jets.

Con las seis respuestas de la BDT se hizo el emparejamiento de jets para averiguar las eficiencias. Se han hallado las eficiencias de las seis BDTs para cada categoría y a mayores se ha hallado la eficiencia del *Kinfitter* para poder comparar. Se ha comprobado la eficiencia del *Kinfitter* cuyos porcentajes de emparejamiento han sido superiores a los de las BDT en todos los canales. Este método es el método canónico utilizado por el CMS así que este resultado cae dentro de lo previsto. Para las BDT, las eficiencias estaban bastante por debajo del *Kinfitter*. Las BDT con mejores resultado fueron las que combinaban fondos, los cortes en la masa invariante no parecieron afectar mucho al resultado final y el corte en la SumCSV tampoco mejoró la eficiencia. En donde hay porcentajes de eficiencia altos es al reconocer al menos un jet, esta categoría, en general, no comparable con la reconstrucción cinemática pero es posible hacer una segunda iteración sobre los parámetros de la primera BDT para intentar hallar el segundo jet. Las eficiencias de identificación de jets adicionales en las BDT han alcanzado máximos de hasta el 57 % mientras que el *Kinfitter* ha alcanzado máximos del 75 %. Con esto se concluye que el método más eficiente al la hora de identificar jets adicionales es el *Kinfitter* pero la BDT puede ser una alternativa en algún caso si no se dispone de tanto poder computacional.

La necesidad de encontrar métodos alternativos de identificación de jets ha hecho de este un proyecto interesante. La BDT no se ha mostrado como el método más eficiente a la hora de emparejar jets pero si se puede emplear como método alternativo a la reconstrucción cinemática por su ejecución relativamente rápida.

Como nota personal y ya para finalizar, este proyecto me ha parecido muy interesante sobretodo por las herramientas y habilidades que he adquirido por el camino. Me he familiarizado con entornos virtuales, linux, C++ y machine learning que son herramientas que probablemente me sean muy útiles en el futuro y me abran muchas puertas.

Para comenzar este TFG primero se tuvo que superar una barrea técnica. ROOT y todas las herramientas que se han utilizado están en C++, lenguaje que no se estudia en la carrera, así que se ha tenido que empezar a estudiar lo básico de este lenguaje a priori del comienzo de este trabajo.

También se ha tenido que trabajar en entornos virtuales de Linux, en Ubuntu, más concretamente, se ha aprendido a navegar por las terminales, como instalar una máquina virtual, a depurar código y a localizar errores. El *machine learning* goza de popularidad en el mundo

del análisis de datos por lo que haber trabajado y poseer ya experiencia resulta muy útil puesto que esta rama de la ciencia va a seguir creciendo.

Referencias

- D. H. Perkins, An Introduction to High Energy Physics. Cambridge, UK: Cambridge University Press, 4th ed., 2000.
- [2] B. R. Martin, Nuclear and Particle Physics: An Introduction. Hoboken, NJ: Wiley, 2nd ed., 2006.
- [3] M. L. Perl and et al., "Evidence for anomalous lepton production in $e^+ e^-$ annihilation," *Physical Review Letters*, vol. 35, p. 1489, 1975.
- [4] F. Abe and et al. (CDF Collaboration), "Observation of top quark production in *pp* collisions with the collider detector at fermilab," *Physical Review Letters*, vol. 74, p. 2626, 1995.
- [5] J. A. Brochero Cifuentes, "Medida de la sección eficaz de producción de pares top anti-top en canal dileptonico a $\sqrt{8}$ TeV con el detector CMS," 2014.
- [6] C. Muñoz Díaz, "Análisis de las propiedades de producción de jets en colisiones protónprotón en el LHC," 2021.
- [7] "CERN european organization for nuclear research." https://www.home.cern/. Accessed on July 4, 2023.
- [8] C. Collaboration, CMS Physics: Technical Design Report Volume I: Detector Performance and Software. Geneva: CERN, 2006.
- [9] "Pagina web de ROOT." https://root.cern/. Accessed on July 1, 2023.
- [10] ATLAS Collaboration, "Measurement of the w boson mass with the atlas detector." https: //atlas.cern/Updates/Briefing/2023-W-Mass-Measurement, 2023. Accessed on July 4, 2023.
- [11] A. Beiser, Concepts of Modern Physics. New York, NY: McGraw-Hill, 6th ed., 2003.
- [12] W. Cottingham and D. Greenwood, An Introduction to the Standard Model of Particle Physics. Cambridge Univ. Press, 2007.
- [13] W. Leo, Techniques for Nuclear and Particle Physics Experiments. Springer Verlag, 2nd ed., 1994.
- [14] B. Roe, Solutions Manual for Particle Physics at the New Millennium. Springer, 6th ed., 1996.