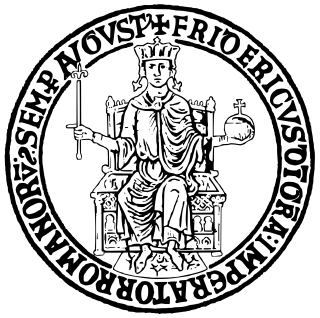
UNIVERSITÀ DEGLI STUDI DI NAPOLI "FEDERICO II"



Scuola Politecnica e delle Scienze di Base
Area Didattica di Scienze Matematiche Fisiche e Naturali
Dipartimento di Fisica "Ettore Pancini"

Laurea Triennale in Fisica

Studio ed ottimizzazione dell'identificazione del quark top per ricerche di nuovi bosoni W' tramite algoritmi di machine learning con il rivelatore CMS ad LHC.

Candidato:
Francesco Montano
Matricola: N85001350
Relatori:
Professor Luca Lista
Dottor Alberto Orso Maria Iorio

Anno Accademico 2020/21

Alla mia famiglia.

Indice

Introduzione

Il modello standard (MS) è la teoria fisica che descrive il mondo sub-nucleare, ovvero la fisica al di sotto della scala del fermi, 10^{-15} m. Dalla sua definitiva formulazione, avvenuta negli anni settanta del secolo scorso, è l'unica teoria esistente che, conciliando la meccanica quantistica e la relatività ristretta, descrive la fisica fondamentale di tre su quattro interazioni. Il MS mostra anche molti limiti, tra i più importanti: la mancata descrizione dell'interazione gravitazionale, l'assenza di candidati di materia oscura ed, inoltre, assume una massa nulla per i neutrini, in contraddizione con le recenti evidenze sperimentali sull'oscillazione di neutrino. E' necessario, quindi, estendere il MS tramite una teoria più generale che sia in grado di fornire soluzioni ai problemi ancora aperti, e di cui il MS stesso risulti essere una rappresentazione alle energia e scale attualmente accessibili. Le ricerche di evidenze di teorie oltre il MS sono un punto centrale per esperimenti come il Compact Muon Solenoid (CMS) in grado di esplorare la fisica nella scala dei TeV, all'acceleratore Large Hadron Collier (LHC).

Tra le possibili indicazioni della presenza di una nuova teoria, c'è l'esistenza di nuovi bosoni W' e Z' con caratteristiche simili a quelle dei bosoni mediatori, W e Z, del MS. L'esperimento CMS sta cercando in particolare il bosone W' che si presume si crei attraverso il processo di annichilazione di due quark nelle collisioni protone-protone ad LHC.

L'obiettivo di questa tesi è di studiare i quark top, provenienti dal decadimento del bosone W' facendo uso di algoritmi di machine learning come il Random Forest, utilizzando nuove tecniche per identificare le variabili da utilizzare, dette anche *feature*. Andremo a trovare quali feature del quark sono più utili alla sua ricostruzione utilizzando metodi standard e un nuovo algoritmo, Boruta, allo scopo di ottimizzare la selezione del quark top con un numero minimo di variabili. L'obiettivo è di aumentare l'efficacia di ricostruzione di quark top provenienti dal decadimento del bosone W' anche per energie molto alte.

Questa tesi è strutturata in 4 capitoli, in cui ognuno descrive un argomento specifico, in ordine: Modello Standard, LHC e CMS, algoritmi di machine learning e l'analisi svolta.

1 Modello Standard

Il modello standard è una teoria fisica sviluppata durante la seconda metà del XX secolo fino alla metà degli anni settanta che descrive le particelle fondamentali, ovvero le componenti indivisibili della materia, e le interazioni che le coinvolgono. [1]

1.1 Particelle elementari

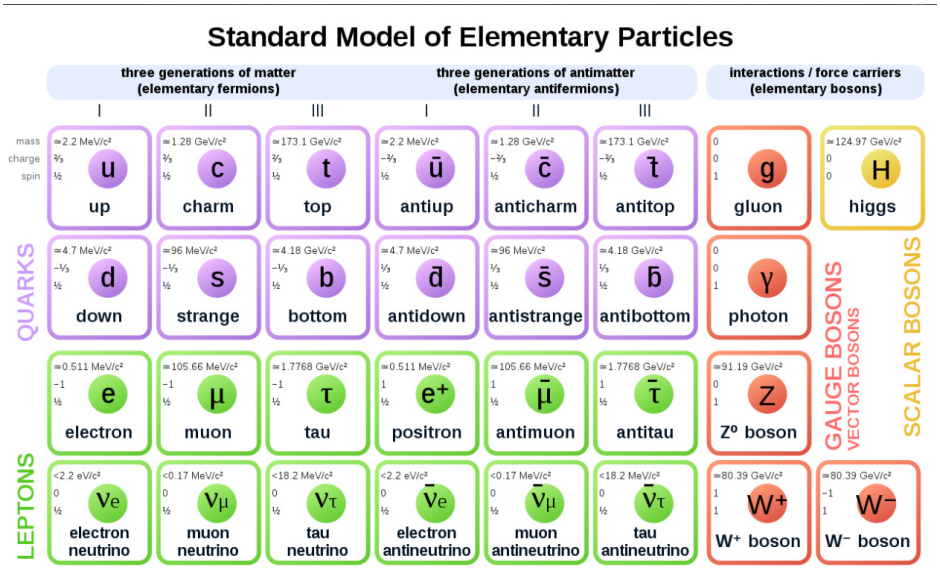


Figura 1: Particelle fondamentali descritte dal modello.

Le particelle descritte possono essere raggruppate in due gruppi:

- **Fermioni**: particelle aventi spin semintero, costituiscono la materia ordinaria dell'universo. Essi vengono divisi in tre famiglie o generazioni. La prima è quella più abbondante nell'universo, che costituisce gran parte della materia ordinaria, mentre le altre due sono più rare e con masse maggiori;
- **Bosoni**: particelle aventi spin intero, sono i mediatori delle interazioni fondamentali. Essi si suddividono in bosoni vettori e bosoni scalari.

1.1.1 Fermioni

I fermioni sono particelle con spin 1/2 e seguono la statistica di Fermi-Dirac. Essi vengono suddivisi in due gruppi: leptoni e quark.

- I **leptnoni** possono essere coinvolti con interazione debole, elettromagnetica e gravitazionale. Possono essere distinti in leptoni carichi o a carica nulla, detti anche neutri. Quelli carichi sono: elettrone (e), muone (μ) e tauone (τ) e hanno tutti carica -1, in unità di carica dell'elettrone e; invece, le rispettive particelle a carica nulla sono i neutrini. Un leptone carico e il suo corrispettivo neutrino costituiscono una famiglia. Si introduce un numero quantico: il numero leptonico di famiglia, $+1e$ per tutti i leptoni e -1e per gli anti-leptoni. In ogni reazione in cui sono coinvolti i leptoni, il loro numero leptonico di famiglia si conserva e tale conservazione non viene infranta da nessuna interazione.

- I **quark** sono i costituenti deli nuclei che formano l'atomo. Essi risentono di tutte le quattro interazioni fondamentali. Sono tutti carichi ma hanno carica frazionaria: up (u), charm (c) e top (t) hanno carica $2/3e$; invece: down (d), strange (s) e bottom (b) hanno carica -$1/3e$. Essi si uniscono per formare gli adroni, che si differenziano in barioni, unione di tre quark, e mesoni, unione di un quark e anti-quark. Ad ogni quark si può associare una carica di colore che può essere: green (G), red (R) e blue (B) o i rispettivi anticolori. E' osservato che gli adroni sono complessivamente a arica di colore nulla, questo fenomeno è chiamato confinamento del colore ed è il motivo per il quale non è possibile osservare cariche di colore isolate. Effetto fenomenologico importante del confinamento dei quark è l'adronizzazione: quando si separano due quark, il campo dei gluoni forma stretti tubi (o stringhe) cromodinamici di carica di colore, che tendono a riportare assieme i quark, come fossero una striscia di gomma elastica; questo fa si che sia più vantagioso in termini energetici unirsi con altri quark per formare un adrone. Per questo motivo, quando i quark vengono prodotti negli acceleratori di particelle, invece di vedere i singoli quark nei rivelatori, si vedono i jet, ovvero getti di varie particelle neutre dal punto di vista della carica di colore (mesoni e barioni) raggruppate insieme. Infine, si introduce il numero quantico barionico, cioè ad ogni barione viene dato il numero quantico +1, e in tutte le interazioni esso si conserva. Il quark top, essendo stato preso in esame per questo studio, verrà discusso più in dettaglio in seguito.

1.1.2 Bosoni

I bosoni possono essere suddivisi in bosoni scalari, che hanno spin nullo, e vettoriali, che hanno spin 1 e seguono, pertanto, entrambi la statistica di Bose-Einstein.

- I **bosoni vettoriali** del MS, o di **gauge**, sono le particelle mediatrici delle forze. Ci sono bosoni senza massa come i gluoni (g), mediatori della forza forte, e i fotoni (γ), mediatori dell'interazione elettromagnetica; i bosoni con massa W e Z sono i mediatori della forza debole, con W che ha carica $\pm 1e$ e Z con carica nulla.

- Il **bosone di Higgs**, unico bosone scalare nella teoria, svolge un ruolo fondamentale conferendo massa alle particelle fondamentali mediante il meccanismo di Brout-Englert-Higgs. Le teorie di gauge, di per sé, non sono in grado di descrivere bosoni vettori dotati di massa, che sono proibiti dalla simmetria e questo contraddirebbe quanto viene osservato sperimentalmente a proposito dei bosoni W e Z. Il meccanismo di rottura spontanea della simmetria di gauge è tuttavia in grado includere anche i bosoni massivi nel MS, introducendo un ulteriore bosone, a sua volta massivo, il bosone di Higgs, per l'appunto. Esso è stato osservato per la prima volta nel 2012 dagli esperimenti ATLAS e CMS, che hanno misurato una massa di (125.35 ± 0.15) Gev. [2]

1.2 Interazioni Fondamentali

Le interazioni vengono descritte, nella fisica dei quanti, tramite lo scambio di mediatori, cioè di bosoni di gauge, che interagiscono con la materia. Le interazioni descritte dal modello sono le seguenti:

- **Elettromagnetica**: interazione fra due particelle con carica elettrica;

- **Debole**: interazione responsabile del meccanismo dei decadimenti radioattivi beta;

- **Forte**: interazione responsabile della coesione dei costituenti dei nuclei.

Da cui si vede una delle problematiche maggiori del modello, ovvero che non descrive l'ultima interazione fondamentale, quella gravitazionale. Di seguito una panoramica delle interazioni:

Tabella 1: Caratteristiche delle interazioni.

FORZA	**MEDIATORE**	**INT. REL.**	**RAGGIO(m)**
Forte	Gluone	10^{38}	10^{-15}
Elettromagnetica	Fotone	10^{36}	∞
Debole	Bosoni W e Z	10^{25}	10^{-18}

1.2.1 Interazione elettromagnetica

L'interazione elettromagnetica si esplica tra due particelle cariche. La teoria che la descrive è l'elettrodinamica quantistica (QED) che include la teoria della relatività ristretta. La particella mediatrice è il fotone (γ), un bosone di massa nulla e ha range infinito. Essa non viola nessuna legge di conservazione, tranne la terza componente di isospin (I_3).

Feynman, per descrivere l'interazione e per calcolare le ampiezze di transizione in un processo di scattering (o di decadimento) e, quindi, le sezioni d'urto e le costanti di decadimento, introdusse dei diagrammi che prendono il suo nome. Feynman ebbe l'idea di visualizzare i calcoli nel MS, che sono di tipo perturbativo, come la probabilità di un processo, con questi diagrammi, detti diagrammi di Feynman, che visualizzano, appunto, un termine della serie perturbativa dell'ampiezza di scattering per un processo definito dagli stati iniziali e finali.

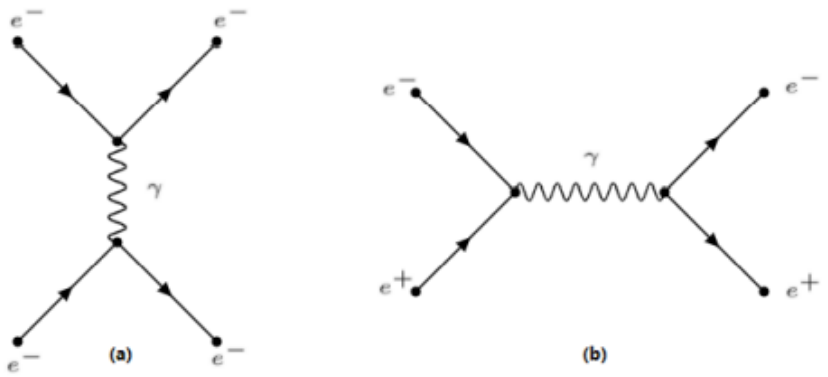


Figura 2: Diagrammi di Feynman in QED (a): scattering elastico elettrone-elettrone; (b): annichilazione elettrone-positrone.

1.2.2 Interazione debole

L'interazione debole, chiamata così perché di intensità minore rispetto alle altre, agisce su ogni particella ed è l'unica, tra quelle descritte dal modello, che agisce sui neutrini. Essa ha due bosoni mediatori: W e Z, che danno luogo a due tipi di interazione deboli: il primo tipo è chiamato "interazione a corrente carica" perché è mediata da particelle dotate di carica elettrica: i bosoni W^+ e W^- , ed è responsabile del decadimento beta; il secondo tipo è chiamato "interazione a corrente neutra" perché è mediata da una particella

neutra: il bosone Z^0ed è responsabile per la rara deflessione dei neutrini. Questi due bosoni sono molto massivi e questo comporta che il range della forza è molto piccolo, sui 10^{-18} m.

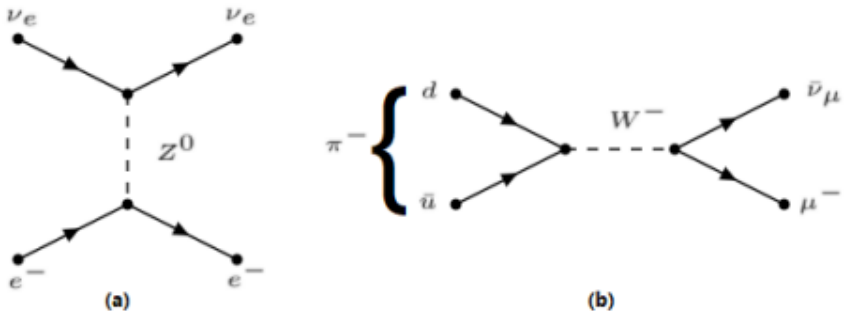


Figura 3: (a): scattering elastico, corrente neutra; (b) decadimento pione, corrente carica.

L'interazione debole è l'unica che non rispetta la conservazione della simmetria di parità P, la proprietà di un fenomeno di ripetersi immutato dopo un'inversione delle coordinate spaziali, della simmetria C, cioè rispetto alla coniugazione di carica e del prodotto CP. Infine, per il teorema CPT, essa non rispetta nemmeno la simmetria T, ovvero per inversione temporale. L'interazione debole è l'unica che può cambiare sapore e famiglia ai quark, attraverso l'interazione a corrente carica e tale fenomeno viene descritto dalla matrice CKM (Cabibbo- Kobayashi-Maskawaed):

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix} \tag{1.1}$$

I quark che partecipano all'interazione debole non sono stati puri di sapore, ma sono invece una combinazione lineare. Gli elementi della matrice CKM rappresentano i coefficienti di questa combinazione lineare, ovvero le ampiezze di transizione dei quark da una famiglia ad un'altra, dove gli elementi nella diagonale principale sono prossimi all'unità.

Ad energie superiori a 100 GeV, l'interazione elettromagnetica e la forza nucleare debole diventano due manifestazioni di un'unica interazione, l'interazione elettrodebole, la cui simmetria è manifesta ad alte energie, ma è rotta spontaneamente, a bassa energia, dal meccanismo di Higgs.

1.2.3 Interazione forte

L'interazione forte, chiamata così perché di intensità maggiore rispetto alle altre forze fondamentali, viene descritta dalla teoria fisica chiamata Cromodinamica Quantistica (QCD). Essa può essere osservata in scala più piccola fra quark costituenti uno stesso nucleone e altre particelle, o in scala più grande fra quark di protoni e neutroni diversi all'interno del nucleo atomico, dove si parla più propriamente di forza nucleare forte o forza forte residua. Nel primo caso le particelle mediatrici dell'interazione sono i gluoni (g), gli unici capaci di trasferire carica di colore; nel secondo di scambiare pioni (π), che sono mesoni e posso avere carica nulla o $\pm 1e$. L'interazione forte rispetta ogni tipo di conservazione.

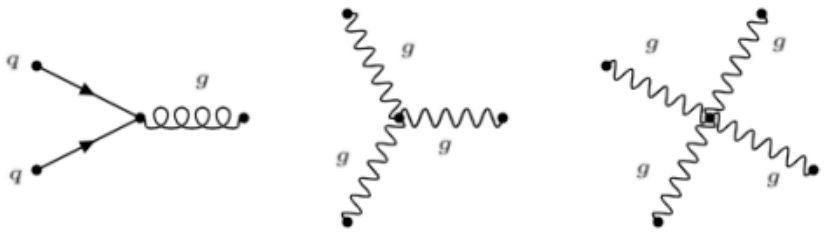


Figura 4: Tipi di diagramma in QCD.

1.3 Quark Top

Il quark top (t) è la particella fondamentale più massiva descritta dal modello e finora osservata. Esso è un fermione della terza generazione con massa m=(173.1±1.4) GeV e carica $+2/3e$. [2] Esso decade solo per interazione debole in un bosone W e in un quark b e il suo tempo di vita media è 2×10^{-25} s, 20 volte più breve della scala di tempo caratteristica per iniziare le adronizzazioni, per cui tipicamente decadono prima di adronizzare. Per questo motivo, il quark t non è mai stato osservato in adroni e perciò può essere studiato come un quark "nudo"; invece, tutti gli altri quark adronizzano e possono essere osservati unicamente all'interno di adroni.

1.3.1 Processi di produzione

Considerata l'elevata massa del quark t, occorrono energie molto grandi per produrlo e, attualmente, l'unica macchina in funzione in grado di raggiungere

tali energie è l'LHC. Come detto prima, si può studiare un quark top "nudo" che si può ottenere nei seguenti processi:

- **canale s**: un quark, annichilendo con un antiquark, produce, mediante lo scambio di un bosone W, un top e un antibottom;
- **canale t**: un quark bottom produce un top scambiando un bosone W con un altro quark;
- **produzione tW**: un quark bottom interagisce con un gluone producendo un quark top ed un bosone W.
- **produzione $t\bar{t}$**: un quark, annichilendo con un antiquark, produce, mediante lo scambio di un gluone, un top e un antitop;

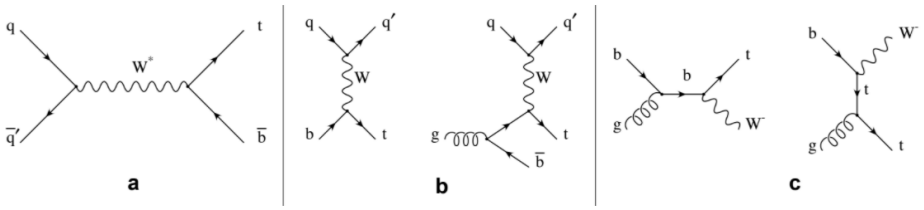


Figura 5: Diagrammi di Feynman per la produzione di un singolo quark top nel canale s (a), nel canale t (b), nella produzione tW.

1.3.2 Processi di decadimento

Come accennato prima, il quark top decade per interazione debole a corrente carica e può decadere in un quark d, s o b, ma nella quasi totalità dei casi esso decade in b, t→Wb. Questo dipende dal fatto che le ampiezze di probabilità, descritta dalla matrice CKM, $V_{qq'}$ siano diverse e in questo caso $V_{tb} >> V_{td}$,V_{ts}. Ciò implica che in questo canale la frazione di decadimento:

$$R = \frac{BR(t \to Wb)}{BR(t \to Wq')} = \frac{V_{tb}^2}{V_{td}^2 + V_{ts}^2 + V_{tb}^2} \tag{1.2}$$

sia prossima ad 1. Successivamente il bosone W decade o in coppie leptone-neutrone o in coppie di quark.

1.4 Oltre il Modello Standard

Come detto in precedenza, il modello standard presenta delle lacune che devono essere colmate con nuove teorie, una di queste è la Teoria della Grande unificazione. Tre delle quattro diverse forze fondamentali, gravità esclusa, hanno intensità comparabili se si raggiunge una scala di energia sufficientemente elevata e potrebbero essere descritte da un'unica teoria, chiamata Grand Unification Theory (GUT), con soglia di energia pari a 10^{16} GeV. Un ulteriore modello, che inglobi anche la gravità, è chiamato Theories 0f Everything (TOE) ed ha una scala di energia prevista a 10^{19} GeV.

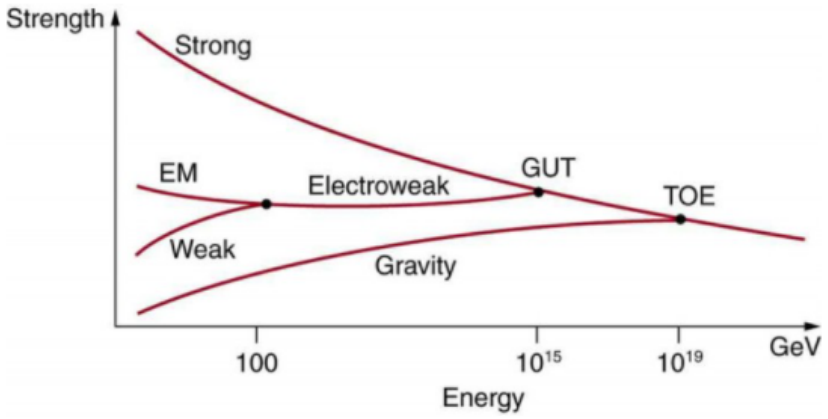


Figura 6: Schema dell'unificazione delle forze.

1.4.1 Bosone W'

Il bosone W', particella di spin 1 e carica $\pm 1e$ che non è stata ancora scoperta in nessun esperimento, è alla base di molte teorie come: left-right symmetric model e nei modelli little Higgs. Si tratta di un nuovo mediatore di una nuova corrente carica che può avere una massa tale da decadere in un quark top e un quark b, precisamente: $W'^{+} \to t\bar{b}$ e $W'^{-} \to \bar{t}b$.[3] Recentemente, negli esperimenti BaBar, Belle e LHCb sono stati misurati decadimenti semileptonici di mesoni B in mesoni D e D* e sono state trovate discrepanze significative con il Modello Standard. La prima discrepanza significativa è nel branching ratio, che indica la frazione di particelle che decadono seguendo quel particolare canale di decadimento rispetto al numero totale di particelle

che decade in qualsiasi canale:

$$R(D) = \frac{\lambda(\overline{B} \to D\tau^-\overline{\nu}_\tau)}{\lambda(\overline{B} \to Dl^-\overline{\nu}_l)} \tag{1.3}$$

$$R(D^*) = \frac{\lambda(\overline{B} \to D^*\tau^-\overline{\nu}_\tau)}{\lambda(\overline{B} \to D^*l^-\overline{\nu}_l)} \tag{1.4}$$

dove l = e, μ. Le misure hanno trovato un eccesso di decadimenti in disaccordo con i valori calcolati dal modello stantard di circa 4 deviazioni standard σ. Il processo alla base è il decadimento via corrente carica $b \to c\tau\nu$, che è fortemente soppresso nel Modello Standard. Un bosone W' che si accoppia con la seconda e terza generazione di fermioni apporterebbe un nuovo contributo fisico al decadimento in grado di spiegare la discrepanza. [4]

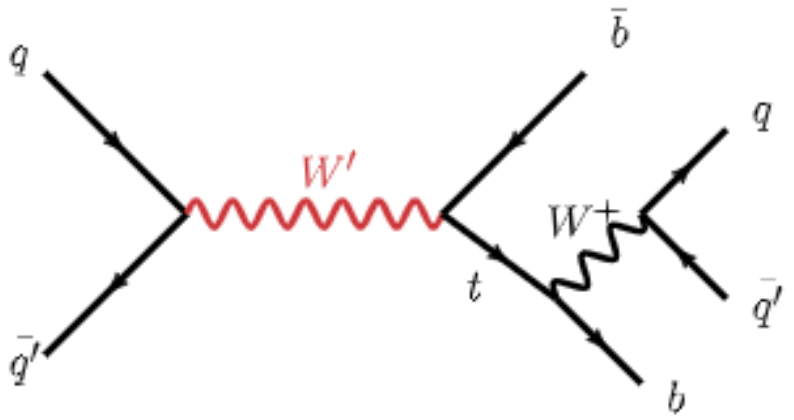


Figura 7: Diagramma di Feynman del decadimento di W'.

2 LHC e CMS

Il Large Hadron Collider (LHC) è un acceleratore di particelle situato presso il CERN (European Organization for Nuclear Research) di Ginevra, utilizzato per ricerche sperimentali nel campo della fisica delle particelle. [5] Il programma scientifico di LHC prevede sette esperimenti. I due esperimenti più grandi sono ATLAS (A Toroidal LHC Apparatus) e CMS (Compact Muon Solenoid) che sono rivelatori di enormi dimensioni e avanzata tecnologia realizzati da collaborazioni internazionali comprendenti oltre 2.000 fisici.

L'obiettivo primario dell'utilizzo dell'LHC è di cercare evidenze scientifiche che possano portare ad un nuovo modello teorico che estenda il Modello Standard e, al contempo, trovi una soluzione a una o a più porblematiche restate aparte.

2.1 Caratteristiche dell'LHC

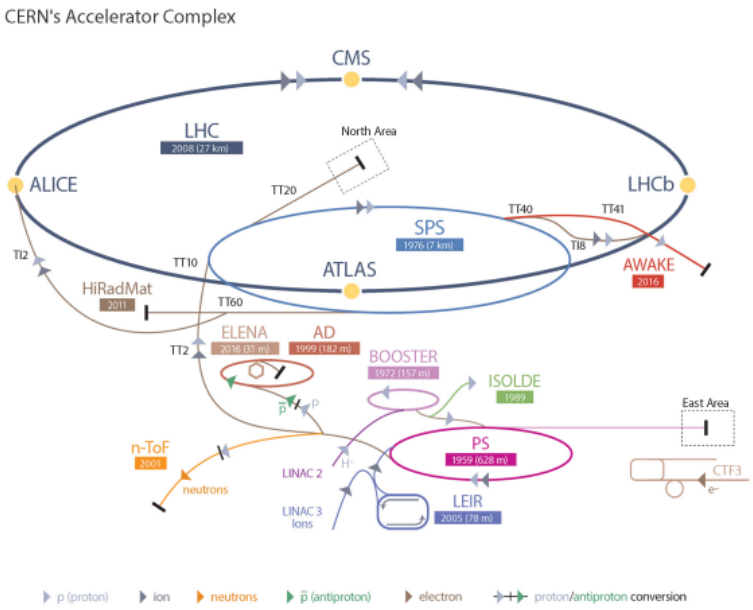


Figura 8: Complesso di acceleratori al Cern.

LHC è l'acceleratore di particelle più grande e potente esistente sulla Terra. Si tratta di un acceleratore di adroni con un'energia nel centro di massa di progetto di circa 14 TeV, costruito all'interno di un tunnel sotterraneo con una circonferenza di circa 27 km, a circa 100 m di profondità. Si trova

nello stesso tunnel realizzato in precedenza per l'acceleratore LEP (Large Electron-Positron Collider). Lo stesso LHC in futuro potrebbe essere usato come iniettore per una macchina ancora più grande: il Future Circolar Collider (FCC), un acceleratore da 100 km di circonferenza.

Le particelle principalmente utilizzate sono protoni provenienti da gas di idrogeno, oppure, in alcuni casi, da ioni di piombo. Mediante l'utilizzo di un campo elettrico si provvede a privare gli atomi di idrogeno dei loro elettroni. I protoni vengono poi accelerati, mediante cavità a radiofrequenza, passando in una serie di acceleratori lineari e circolari: (Linac2), Proton Synchrotron Booster (PSB), Proton Synchrotron (PS) e il Super Proton Synchrotron (SPS) ed, infine, immessi nei due anelli viaggiando in direzioni opposte. I fasci vengono fatti scontrare in quattro punti in prossimità dei maggiori esperimenti. I quattro principali rivelatori di particelle sono:

- **Alice (A Large Ion Collider Experiment)**, che ha lo scopo di studiare urti tra nuclei di piombo a un'energia del centro di massa di 2.76 TeV per coppia di nucleoni. Si prevede che la temperatura e la densità di energia risultanti siano abbastanza per produrre una fase di materia chiamata plasma di quark e gluoni (QGP), nel quale i quark e i gluoni sono liberi;

- **Atlas (A Toroidal LHC ApparatuS)**, progettato per rilevare particelle pesanti a bassa energia, misurando il più ampio intervallo possibile di segnali. Questo per assicurare che, qualunque caratteristica un nuovo processo fisico o una nuova particella possa avere, ATLAS sia in grado di rivelarli e misurare le loro proprietà;

- **CMS (Compact Muon Solenoid),** che verrà descritto in dettaglio nelle prossime pagine.

- **LHCb**, che studia le collisioni di protoni prodotte dall'acceleratore ad energie tra i 7 e i 13 TeV ed ha lo scopo di misurare i parametri della violazione della simmetria di Carica-Parità (CP) e i decadimenti e fenomeni rari relativi agli adroni in cui è presente il quark b, da cui il nome dell'esperimento;

Affinchè i fasci non sfuggano dall'anello, ma rimangano confinati in esso e ben focalizzati, sono disposti, lungo LHC, dei magneti superconduttori: i dipoli curvano il fascio lungo l'anello, i quadrupoli, invece, mantengono il fascio ben focalizzato. I dati raccolti dai rivelatori vengono poi inviati al centro di calcolo. Lo scopo di LHC è quello di rivelare la fisica oltre il Modello Standard con energie di collisione nel centro di massa fino a 14 TeV. Il numero di eventi

al secondo generati nelle collisioni di LHC è dato da:

$$N_{event} = \mathcal{L}\sigma_{event}. \tag{2.1}$$

dove σ_{event} è la sezione d'urto per l'evento in esame e $\mathcal{L}$ la luminosità della macchina. La luminosità della macchina dipende solo dai parametri del raggio e può essere scritta per un fascio che segue una distribuzione gaussiana come:

$$\mathcal{L} = \frac{N_b^2 n_b f_{rev} \gamma_{rev}}{4\pi\epsilon_n \beta^*} F. \tag{2.2}$$

dove N_b è il numero di particelle per bunch, n_b il numero di bunch per fascio, f_{rev} la frequenza di rivoluzione, γ_{rev} il fattore gamma relativistico, n l'emittanza del raggio trasversale normalizzata, β^* la funzione beta nel punto di collisione, e F il fattore geometrico di riduzione della luminosità dovuto al angolo di incrocio nel punto di interazione. [6]

2.2 Esperimento CMS

Il Compact Muon Solenoid, CMS, è un esperimento "general purpose" con lo scopo di studiare la fisica del MS e di cercare indicazione di nuova fisica. L'apparato sperimentale misura 21.6 m di lunghezza per 14.6 m di diametro per un peso totale di circa 12500 t. [7] E' composto da un rivelatore di forma cilindrica, la cui struttura è basata su un magnete solenoidale costituito da una bobina superconduttrice che genera un campo magnetico interno di 3.8 T. Esso è compatto ed ermetico, il che vuol dire che i rivelatori sono disposti in modo da coprire tutto l'angolo solido intorno al punto di interazione che, grazie ai suoi molteplici sottorivelatori, permette di rivelare elettroni, fotoni, muoni ed adroni carichi e neutri. La presenza di neutrini o altre particelle non interagenti è messa in evidenza attraverso l'energia mancante, misurata dallo sbilanciamento dei depositi energetici nel piano trasverso ai fasci.

2.2.1 Sistema di coordinate

La cinematica degli oggetti ricostruiti, utilizzati nell'analisi dei dati di CMS, è descritta tramite un sistema di coordinate cartesiano. Il sistema di riferimento, centrato nel punto di interazione nominale dei fasci, è orientato in modo che l'asse z coincida con la direzione della velocità dei protoni nel fascio, l'asse x diretto verso il centro di LHC e l'asse y diretto verso l'alto. Tuttavia, per la simmetria dell'apparato, utilizziamo coordinate sferiche (r, θ, ϕ), dove ϕ descrive l'angolo sul piano xy mentre θ descrive l'angolo sul piano yz. Una quantità importante legata al prodotto di una collisione ad alta energia è la

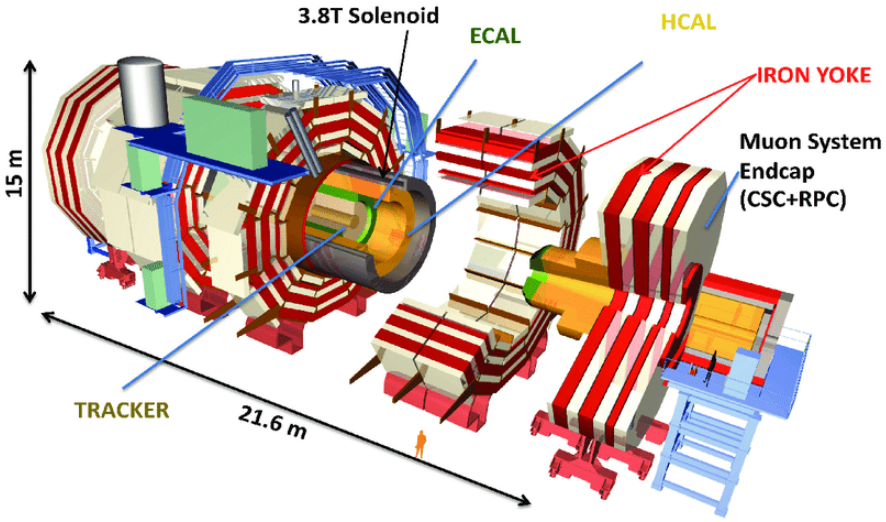


Figura 9: Rappresentazione CMS.

rapidità, che spesso è usata come coordinata alternativa rispetto all'angolo θ del prodotto di una collisione:

$$y = \frac{1}{2} \ln \frac{E + p_z c}{E - p_z c}. \tag{2.3}$$

dove E è l'energia, p_z la quantità di moto lungo la direzione del fascio e c la velocità della luce nel vuoto. Più semplice definire la pseudorapidità, una coordinata spaziale:

$$\eta = -\ln \tan \frac{\theta}{2}. \tag{2.4}$$

dove θ è proprio l'angolo tra il momento della particella prodotto e il fascio. Altre quantità importanti sono il momento trasverso e l'energia trasversa:

$$\vec{p_t} = \sqrt{\vec{p_x}^2 + \vec{p_y}^2}. \tag{2.5}$$

$$E_t = E \sin \theta \tag{2.6}$$

dove p_x e p_y sono le proiezioni del momento della particella lungo l'asse x e y e E è la sua energia. In queste coordinate la distanza angolare ΔR tra due oggetti è un invariante per boost lungo l'asse z:

$$\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}. \tag{2.7}$$

2.2.2 I sottorivelatori

CMS è composto da un magnete solenoidale superconduttore che fornisce un campo magnetico di 3.8 Tesla. [8] All'interno del solenoide sono installati i seguenti rivelatori, partendo dall'interno verso l'esterno:

- **Il Silicion Tracker** è il più vicino al punto di interazione (IP), il punto di collisione tra i fasci di protoni. È, quindi, attraversato da un numero molto grande di particelle, circa 10 milioni di particelle per centimetro quadrato al secondo. Situato al centro del rivelatore, ricostruisce le traiettorie di elettroni, muoni e adroni ad alta energia caricati elettricamente e permette di ricostruire il punto di decadimento di particelle di vita molto breve, come le particelle contenenti quark b. Esso è diviso in due sottorivelatori: il Pixel Tracker è costituito da strati concentrici as anelli di 1.800 piccoli moduli di silicio. La minuscola dimensione dei pixel (100x150) μm^2 consente una misurazione accurata della traiettoria di una particella, che passa attraverso il rivelatore e della sua origine, con una precisione di circa 10 μm per misure trasversali e 10 μm per quelle radiali; il Microstrip Tracker, più esterno, constituito da 4 strati di strip al silicio parallele al fascio, esso ha una risoluzione piu bassa, tra i 35 μm e i 52 μm in direzione radiale, e i 530 μm in direzione trasversale. Entrambi funzionano in modo simile: attraversati da particelle cariche il materiale di cui sono fatti, silicio drogato, viene eccitato in modo da formare coppie e^--lacuna. Gli elettroni, per effetto di un campo elettrico applicato, si muovono verso particolari sensori e danno luogo ad un impulso elettrico che dura un nanosecondo. Il segnale viene poi amplificato e permette di ricostruire la traiettoria delle particelle punto per punto.

- **Il Calorimetro Elettromagnetico (ECAL)** permette di rivelare fotoni ed elettroni e misurarne l'energia con grandissima precisione. Copre la regione con $|n| < 2.5$ e $r < 1.2$ m. Esso è composto di circa 75000 cristalli di tungstato di piombo. Una volta attraversato dalle particelle, il materiale del cristallo produce luce visibile che viene opportunamente raccolta e trasformata in segnale elettrico tramite un foto-diodo a valanga. Questo dispositivo, sviluppato specificamente per essere accoppiato ai cristalli di PWO, viene incollato ad una estremità del cristallo ed è in grado di convertire la luce in segnale elettrico moltiplicandolo immediatamente all'uscita dal cristallo stesso.

- **Il calorimetro adronico (HCAL)** copre la regione con $|n| < 5$, misura l'energia degli adroni, particelle fatte di quark e gluoni, inoltre

fornisce una misura indiretta della presenza di particelle neutre che non interagiscono, come i neutrini. HCAL è un calorimetro a campionamento, che significa che è costruito alternando strati di un denso materiale "assorbitore" con strati di "scintillatore" fluorescente, materiale che produce un rapido impulso di luce quando una particella lo attraversa. Ogni strato corrispondente di assorbitore e di scintillatore è suddiviso geometricamente in mattonelle delle stesse dimensioni. Fibre ottiche speciali raccolgono la luce di scintillazione e la trasmettono a scatole di lettura dove fotomoltiplicatori amplificano il segnale. Quando la quantità di luce in una particolare regione del rivelatore viene sommata su varie mattonelle in profondità, definita come una torre, questo somma di luce fornisce una misura dell'energia della particella.

- **Il magnete**, un solenoide composto da spire di bobina conduttrice che producono un campo magnetico uniforme di 3.8 T. Il magnete è raffreddato ad una temperatura di -268.5º C, permettendo alla corrente elettrica di scorrere senza resistenza nelle sue bobine. Il suo compito è quello di deviare le traiettorie di particelle cariche che provengono dalle collisioni protone-protone di LHC. Maggiore è l'impulso di una particella, minore è la sua deviazione in campo magnetico, e quindi ricostruire la sua traiettoria ci permette di ottenere una misura del suo impulso.

- **Il sistema a muoni** ha il compito di rilevare i muoni e riveste un ruolo di grande importanza come suggerisce il nome dell'esperimento. Siccome i muoni possono penetrare parecchi metri di materiale senza interagire, al contrario della maggior parte delle particelle prodotte alle collisioni di LHC, i muoni non sono "fermati" da nessuno dei calorimetri di CMS. I rivelatori di muoni sono quindi collocati all'esterno del magnete, dove solo i muoni sono in grado di giungere e lasciare un segnale. Il rivelatore di muoni è suddiviso in quattro "stazioni" a distanza crescente dal punto di interazione. Ogni stazione è composta da svariate camere individuali. Ogni camera è a sua volta costituita di vari strati indipendenti. Un muone viene misurato dalla traccia curva formata dall'interpolazione dei segnali nelle quattro stazioni. Sfruttando le informazioni sulla posizione misurata dalle camere a muoni con la posizione misurata dal tracciatore centrale si ricava con grande precisione la traiettoria di un muone. Dalla curvatura della traiettoria nel campo magnetico possiamo misurare l'impulso del muone. In totale ci sono 1400 camere a muoni: 250 camere a deriva "Drift Tubes" (DT) e 540 camere proporzionali con catodo segmentato in strips "Cathode

Strip Chambers" (CSC) tracciano le particelle che le attraversano e forniscono un segnale di trigger, mentre 610 camere resistive "Resistive Plate Chambers" (RPC) formano un sistema di trigger ridondante, che permette di decidere rapidamente se accettare o meno un muone.

- **I Gas Electron Multiplier (GEM)** sono rivelatori di recente aggiunta per l'upgrade di CMS nella parte a basso angolo del rilevatore. Sono costituiti da tre strati, ciascuno dei quali è formato da una lamina di poliimmide rivestita di rame spessa 50 μm. Queste camere sono riempite con una miscela di gas Ar/CO2, dove si verificherà la ionizzazione primaria dovuta ai muoni incidenti. Le camere GEM forniranno ridondanza e punti di misurazione aggiuntivi, consentendo una migliore identificazione della traccia dei muoni e anche una copertura più ampia nella regione più avanzata.

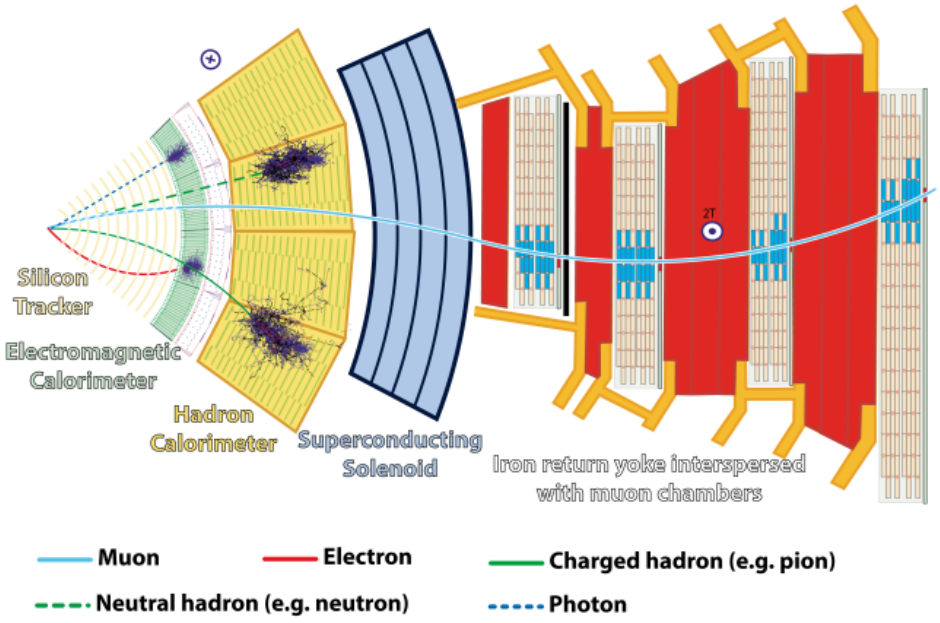


Figura 10: Sezione trasversale del CMS.

Gli eventi di interesse sono selezionati utilizzando un sistema di trigger a due livelli. Il primo livello (L1), composto da processori hardware personalizzati, utilizza le informazioni dei calorimetri e rilevatori di muoni per selezionare eventi a una velocità di circa 100 kHz in un intervallo di tempo inferiore di 4 µs. Il secondo livello, noto come trigger di alto livello (HLT), è costituito da una fattoria di processori che eseguono una versione del software di ricostruzione completa degli eventi ottimizzata per la veloce elaborazione e riduce la

frequenza degli eventi a meno di 1 kHz prima della memorizzazione dei dati. Ogni singola particella viene ricostruita ed identificata tramite l'algoritmo Particle Flow che ottimizza tutte le informazione ricevute da ogni elemento del CMS. [9]

3 Algoritmi di Machine Learning

Con il termine Machine Learning (ML) ci si riferisce all'apprendimento automatico inteso come abilità delle macchine di apprendere senza essere state esplicitamente e preventivamente programmate. [10] Gli algoritmi ML permettono ai computer di imparare, mediante un allenamento su un campione di dati di *input* e usare analisi statistiche per emettere un informazione di *output* sul campione stesso, come: classificazione dei dati rispetto a categorie, stabilire relazioni tra variabili, etc... Per questo motivo, il ML aiuta i computer a costruire dei modelli a partire da esempi di dati, in modo da poter realizzare un sistema automatico di decisioni basato sugli input ricevuti.

3.1 Allenamento degli algoritmi

Gli algoritmi di machine learning devono essere allenati con degli esempi che sono forniti dallo sviluppatore e i metodi usati sono:

- **l'apprendimento supervisionato** che allena un algoritmo basato su dati di input e output classificati, in gergo "etichettati", dagli umani. Vengono date in input variabili indipendenti x_1, x_2,...,x_n e variabili obiettivo, *target* y_1,y_2,...,y_n da cui si genera la funzione $f(x_1, x_2, \ldots, x_n) = y$. A partire dal campione di allenamento, la macchina implementerà un modello capace di prevedere autonomamente il risultato y_1,y_2,...,y_n a partire da un nuovo set dati di x_1, x_2,...,x_n.

- **l'apprendimento non supervisionato** in cui il modello ha lo scopo di trovare una struttura negli input forniti, senza che gli input vengano etichettati in alcun modo.

- **apprendimento per rinforzo** in cui il modello interagisce con un ambiente dinamico nel quale cerca di raggiungere un obiettivo, avendo un input che gli dice solo se ha raggiunto l'obiettivo. [11]

In generale il set di dati, usato per allenare un algoritmo di machine learning, viene diviso in un *training set*, un insieme di dati che vengono utilizzati per addestrare un sistema supervisionato, e in un *test set*, con il quale il sistema addestrato viene messo alla prova. Un modello può essere soggetto a due errori comuni: l'overfitting e l'underfitting.

- **Overfitting** è un problema che si presenta quando il fit crea un modello molto specifico al campione di allenamento, e di fatto inizia a imparare le fluttuazioni statistiche del set. In questi casi, il fit introduce dei paremetri in più rispetto a quelli necessari per descrivere il

fenomeno. Risulterà, quindi, un modello troppo sensibile ai training set, e dunque poco generalizzabile ad altri altri casi;

- **Underfitting** è un problema di apprendimento che si verifica quando si ci basa su pochi parametri. In questi casi, ha prestazioni scarse sui training set e si riconduce ad un modello troppo semplice che non riesce a predire nuovi risultati.

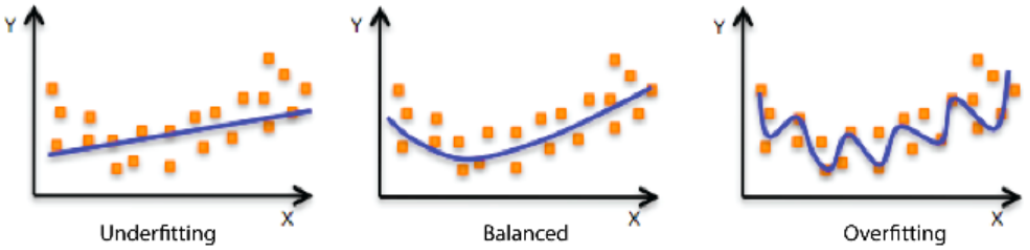


Figura 11: Esempi di overfitting e underfitting.

Una categorizzazione dei compiti dell'apprendimento automatico si può effettuare in base al tipo di output prodotto dal sistema:

- nella **classificazione** i dati di output sono divisi in due o più classi e il sistema di apprendimento deve produrre un modello che assegni gli input non ancora visti a una o più di queste.

- Nella **regressione** che è anch'essa un problema supervisionato, l'output e il modello utilizzati sono continui.

- Nel **clustering** un insieme di input viene diviso in gruppi. Diversamente da quanto accade per la classificazione, i gruppi non sono noti prima, rendendolo tipicamente un compito non supervisionato.

Per avere dei risultati a maggior potere predittivo si utilizza l'apprendimento di insieme, un processo di addestramento di più modelli e di combinazione dei loro risultati. I diversi modelli vengono utilizzati come base per creare un modello predittivo ottimale, riducendo varie problematiche come l'overfitting in taluni casi. Solitamente raggiungono performances più elevate se rapportate ai classificatori singoli. Naturalmente richiedono molto più tempo di addestramento, in quanto al posto di un solo classificatore devono essere addestrati centinaia o migliaia di classificatori. Possono essere di tre tipi:

- **Bagging** o **bootstrap aggregating**, dove più modelli dello stesso tipo vengono addestrati su dataset diversi, ciascuno ottenuto dal dataset

iniziale tramite campionamento casuale con rimpiazzo. Quindi, si generano diversi training set con il bootstrap, campionamento casuale, e le previsioni di ogni modello vengono aggregate per combinarle in modo tale che la previsione finale consideri tutti i risultati possibili. L'aggregazione può essere effettuata in base al numero totale di risultati o alla probabilità di previsioni derivate dal bootstrap di ogni modello nella procedura.

- **Boosting**, che a differenza del bagging, ciascun classificatore influisce sulla votazione finale con un certo peso. Tale peso sarà calcolato in base all'errore di accuratezza che ciascun modello commetterà in fase di learning.

- **Stacking**, mentre nel bagging l'output era il risultato di una votazione, nello stacking viene introdotto un ulteriore classificatore (detto meta-classificatore) che utilizza le predizioni di altri sotto-modelli per effettuare un ulteriore learning. [12]

3.1.1 Valutazione dei classificatori

Il risultato predetto da un classificatore rientra in una delle tipologie riportate nella seguente figura:

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Figura 12: Possibili predizioni di un evento

dove sulle righe gli eventi vengono classificati secondo il modello, e sulle colonne vengono inserite le classi effettive. In colore verde indichiamo un evento che è stato classificato in maniere corretta, che nel testo abbiamo indicato con la notazione "True Positive (TP)" e "True Negative (TN)"; mentre con quelli in rosso indichiamo errori di predizione, che abbiamo chiamato "False

Positive (FP)" e "False Negative (TN)". Per la valutazione dell'algortimo introduciamo tre parametri: l'accuracy, il recall e la precision.

- **Accuracy**: indica in percentuale l'accuratezza dell'algoritmo, definita come il numero di classificazioni corrette dell'algoritmo sul totale dei casi:

$$Accuracy = \frac{Numero\ di\ predizioni\ corrette}{Numero\ totale\ di\ predizioni} = \frac{TP + TN}{TP + FP + FN + TN}. \tag{3.1}$$

- **Precision**: misura la precisione del modello e consiste nel rapposto tra il numero di predizioni corrette sul totale di casi predetti.

$$Precision = \frac{True\ Positive}{Actual\ results} = \frac{TP}{TP + FP}. \tag{3.2}$$

  Quando un modello è preciso per una classe, ogni volta che prevede l'evento sbaglia raramente.

- **Recall**: misura la sensibilità del modello. E' il rapporto tra le previsioni corrette per una classe sul totale dei casi corretti.

$$Recall = \frac{True\ Positive}{Predicted\ Results} = \frac{TP}{TP + FN}. \tag{3.3}$$

3.2 Decision Tree

Un tipo molto importante di algoritmo di machine learning è il decision tree. [13] Rappresentato come in figura 13, esso è un modello predittivo ed hacome punto di partenza un singolo nodo (radice dell'albero), che rappresentauna variabile; il decision tree pone una domanda a questa variabile da cui pu'oottenere delle risposte, cio'e delle ramificazioni possibili (rami dell'albero), eda cui si giunge ad nodo interno, che rappresenta il possibile valore di quellavariabile, che pu'o avere, a sua volta, altre ramificazione fino a giungere airisultati finali (foglie dell'albero).

Un decision tree viene utilizzato per classificare le istanze di grandi quantità di dati. In questo ambito un decision tree descrive una struttura ad albero dove i nodi foglia rappresentano le classificazioni e le ramificazioni l'insieme delle proprietà che portano a quelle classificazioni. Di conseguenza ogni nodo interno risulta essere una macro-classe costituita dall'unione delle classi associate ai suoi nodi figli. In molte situazioni è utile definire un criterio di arresto (halting), o anche criterio di potatura (pruning) al fine di determinarne la profondità massima. Questo perché il crescere della profondità

di un albero (ovvero della sua dimensione) non influisce direttamente sulla bontà del modello. In generale, l'obiettivo è quello di dividere la popolazione iniziale per il valore di una variabile che permette di creare due gruppi il più omogenei possibile internamente e il più disomogenei possibile tra loro, per questo scopo la costruzione dell'albero viene guidata, tipicamente, da due parametri: l'indice di Gini e la variazione dell'entropia.

$$I_G(i) = 1 - \sum_{j=1}^{m} f(i,j)^2. \tag{3.4}$$

L'indice di Gini raggiunge il suo minimo (zero) quando il nodo appartiene ad una singola categoria. Esso rappresenta l'omogeneità del nodo o quanto sia impuro.

$$I_E(i) = -\sum_{j=1}^{m} f(i,j) \log f(i,j). \tag{3.5}$$

In entrambe le formule f rappresenta la frequenza del valore j nel nodo i. Poiché, in generale, in un buon albero di decisione i nodi foglia dovrebbero essere il più possibile puri (ovvero contenere solo istanze di dati che appartengono ad una sola classe), un'ottimizzazione dell'albero consiste nel cercare di minimizzare il livello di entropia man mano che si scende dalla radice verso le foglie. In tal senso, la valutazione dell'entropia determina quali sono, fra le varie scelte a disposizione, le condizioni di split ottimali per l'albero di classificazione.

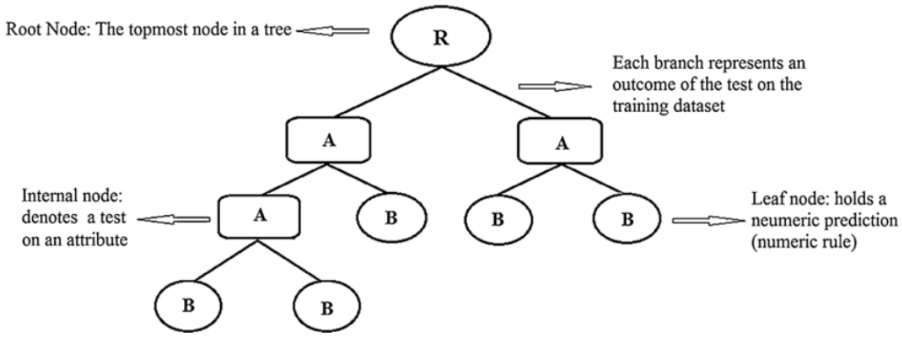


Figura 13: Schema di un Decision Tree.

3.3 Random Forest

Il decision tree ha in sè alcune problematiche: i casi frequenti di overfitting o il fatto che se i dati di addestramento vengono modificati, il decision tree

risultante può essere molto diverso e, di conseguenza, le previsioni possono essere molto diverse. [14] Per superare questi problemi, esiste un'evoluzione del decision tree, il random forest, che può essere inteso come un insieme di decision tree.

Il random forest è un algoritmo di apprendimento supervisionato che utilizza il metodo di apprendimento dell'insieme per la classificazione e la regressione. L'algoritmo di addestramento per le foreste casuali applica la tecnica generale del bagging ai decision tree. [15] Dato un training set $X = x_1$, x_2,...,x_n con risposte $Y = y_1$, y_2, ..., y_n, si ripete il bagging B volte, quindi ho B set training di dimensione n. Per ogni set training si allena un decision tree e si aggregano poi le singole previsioni dei campioni non visti x' facendo una media delle singole.

$$\tilde{f} = \frac{1}{B} \sum_{b=1}^{B} f(x')_b. \tag{3.6}$$

La procedura di bootstrap porta a migliori prestazioni del modello perché riduce la varianza del modello, senza aumentare il bias. Ciò significa che mentre le previsioni di un singolo tree sono altamente sensibili al rumore nel suo set di allenamento, la media di molti alberi non lo è, purché i tree non siano correlati. Il semplice addestramento di molti alberi su un singolo set di addestramento darebbe tree fortemente correlati; il campionamento bootstrap è un modo per de-correlare gli alberi mostrando loro diversi set di allenamento. Inoltre, una stima dell'incertezza della previsione può essere effettuata come deviazione standard delle previsioni da tutti i singoli alberi di regressione su x':

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B} (f(x')_b - \tilde{f})}{B-1}}. \tag{3.7}$$

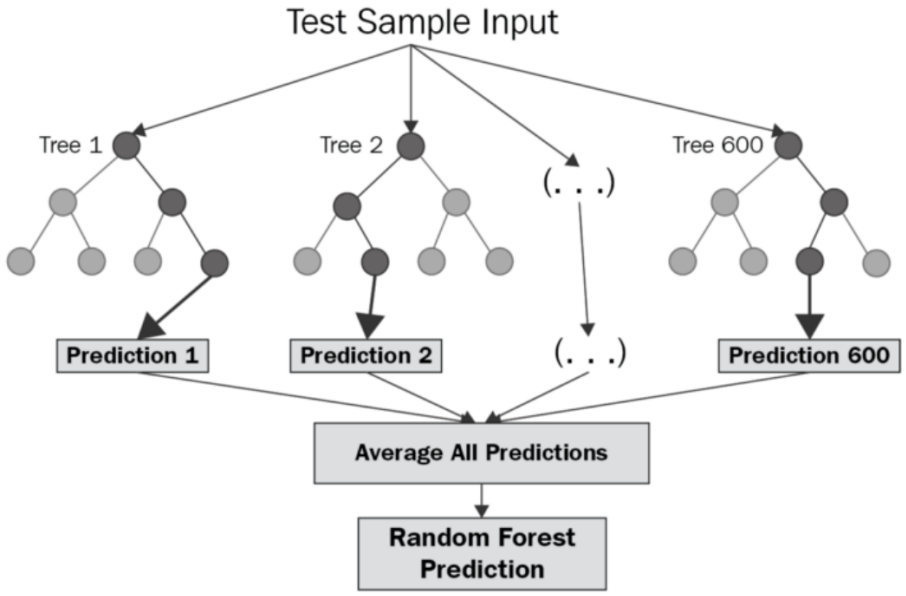


Figura 14: Schema di un Random Forest.

3.3.1 Feature Importance

Uno dei punti cruciali per gli algoritmi di ML è la selezione delle variabili, o *feature*, da utilizzare. Aggiungere feature sovrabbondanti può rendere l'allenamento più lungo senza aggiungere ulteriore informazione, riducendo le performance favorendo l'overfitting. Alcuni algoritmi, come il random forest, consentono di individuare tra le feature analizzate quelle più rilevanti per l'allenamento, la cosidetta *feature importance*. Il random forest trova l'importanza di una singola feature semplicemente calcolando quanto i nodi dell'albero, che utilizzano tale feature, riducono le impurità sulla media di tutti gli alberi decisionali. Si definisce l'importanza del nodo:

$$n_j = w_j C_j - w_{left_j} C_{left_j} - w_{right_j} C_{right_j}. \tag{3.8}$$

con n_j l'importanza del nodo j-esimo, w_j numero ponderato di campioni raggiuti nel nodo j-esimo, C_j il valore di impurità del nodo j-esimo, e con le etichette 'left' e 'right' si denominano i due nodi figli del nodo j-esimo. Il valore di importanza di ogni feature di un albero è dato, quindi:

$$f_i = \frac{\sum_j^{node\ j\ splits\ on\ feature\ i} n_j}{\sum_k^{all\ nodes}}. \tag{3.9}$$

Normalizzandoli:

$$\tilde{f}_i = \frac{f_i}{\sum_j^{all\ nodes} f_j}. \tag{3.10}$$

Le importance features finali sono le medie di tutti gli decision tree:

$$R\tilde{f}_i = \frac{\sum_j^{all\ nodes} \tilde{f}_i}{T}. \tag{3.11}$$

Dove T è il numero dei decision tree.

3.4 Boruta

Boruta è un algoritmo che estende il random forest e ogni algoritmo decisionale, allo scopo di individuare le feature più importanti per l'algoritmo. Il principio di funzionamento si incentra sulla creazione di feature aggiuntive che sono delle copie casuali, chiamate *shadow feature*, delle feature di input, e di addestrare dei classificatori che si basano su questo dataset esteso.

Per comprendere l'importanza delle feature, boruta non le fa competere tra loro, ma le compara a tutte le shadow feature generate, e solo le feature che sono statisticamente più importanti delle shadow feature vengono mantenute, poiché contribuiscono di più alle performance del modello. Quindi, Boruta è pensato esplicitamente, a differenza del random forest, per selezionare le feature più importanti, dando una soglia con cui eliminare le feature. In pratica, a partire da un set di dati X, per ognuna delle feature viene creata una shadow feature ricombinando casualmente i valori di quella feature all'interno del dataset, in modo tale da togliere ogni correlazione con le variabili target y. Si viene così a creare un dataset che ha il doppio delle colonne di quello di partenza e lo chiameremo X_{boruta}. Ora, utilizzando il Random Forest, calcola l'importanza di ciascuna feature, ombra e originali, e le confronta. In questo modo una feature è considerata rilevante se la sua importanza è più alta della importanza massima di tutte le feature ombra. Quando l'importanza di una feature è superiore alla soglia, viene chiamata "hit". Rimuove, quindi, le feature che sono considerate non importanti dal dataset, tra X e X_{boruta}. Infine, ripete la generazione casuale, così il numero di successi ottenuti si compara con una distribuzione binomiale con probabilità al 50% che corrisponde ad una feature su cui non abbiamo informazioni. La soglia di accettazione di una feature è definita come la massima importanza della caratteristica registrata tra le shadow feature e non è rigida tra un'area di rifiuto e un'area di accettazione, e sono individuate tre aree:

- una superficie di rifiuto: le feature che finiscono qui sono considerati come rumore, in modo che vengono eliminati;
- una superficie di irresolution: Boruta è indeciso sulle feature che si trovano in questa zona;

- una superficie di accettazione: le caratteristiche che sono qui sono considerati predittivo, in modo che siano mantenuti.

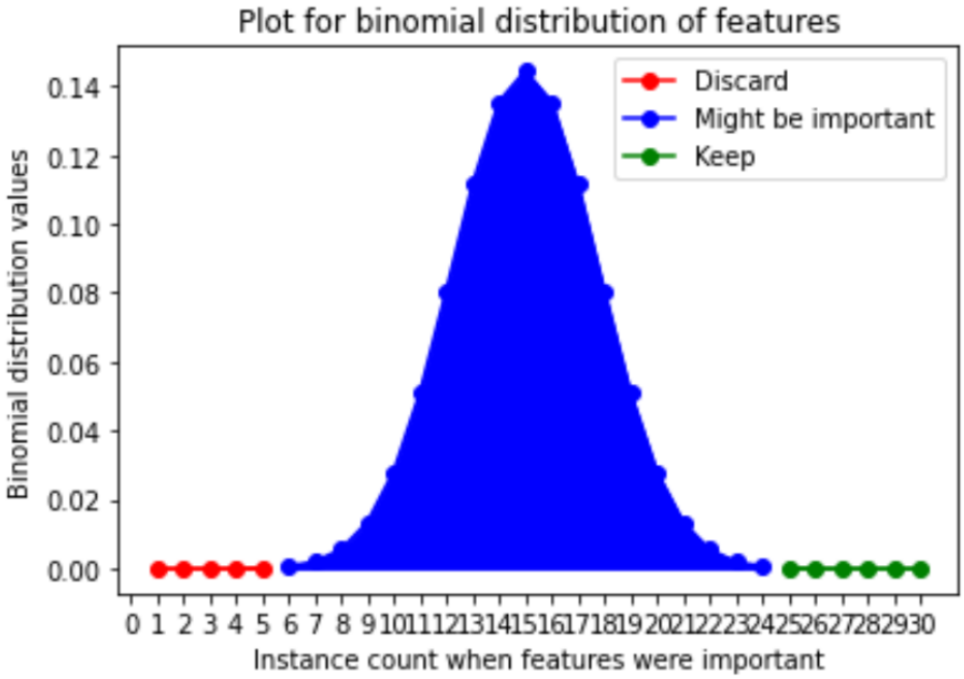


Figura 15: Aree di accettazione di una feature con Boruta.

4 Ricostruzione degli oggetti fisici

Come campione dati presi per lo studio in esame, stiamo usando un campione di dati simulati dell'esperimento CMS ad LHC. Come detto in precedenza, in questo esperimento si studia la collisione di due protoni e il processo considerato è l'ipotetica produzione di un nuovo bosone, W', che decade in un quark top e bottom; il quark t, poi, decade tramite interazione debole, generando un quark b e il bosone W che produrrà leptoni o adroni. Scopo della tesi è di andare a studiare il quark t generato dal decadimento del bosone W' come nella seguente figura. [16] I quark top, non vengono misurati direttamente,

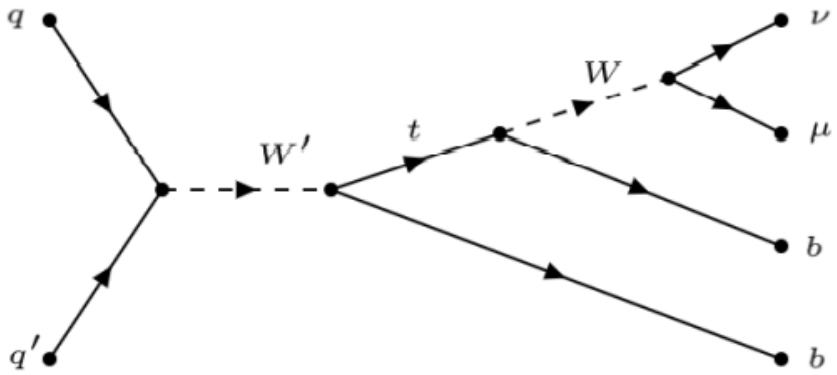


Figura 16: Diagramma di Feynman della produzione del bosone W', della sua catena di decadimento e di quella del quark top risultante.

ma i loro 4-momenti vengono ricostruiti, con un'analisi da parte di CMS, dai suoi prodotti di decadimento e le proprietà di tali prodotti vengono associate al quark top, utilizzando la conservazione del quadrimomento. Passeremo a descrivere ora la ricostruzione delle particelle rivelate all'interno di CMS e le loro proprietà.

4.1 Ricostruzione e descrizione degli oggetti fisici

Gli oggetti visibili all'interno di CMS sono: leptoni, elettroni, muoni, fotoni e jet adronici, ricostruiti mediante l'uso dell'algoritmo Particle Flow, che combina le informazioni ricevuti da tutti i sottorivelatori. [17]

- **Jet**: i quark e i gluoni generati dalle collisioni p-p adronizzano prima di poter interagire direttamente con il rivelatore, ovvero creano sciami di adroni dalla forma conica chiamati jet. I jet vengono ricostruiti

tramite l'algoritmo anti-k_T, al fine di ottenere jet con variabili cinematiche simili a quelle dei quark iniziali. Dopo aver corretto l'energia dei jet in modo da tener conto della risposta del rivelatore, si selezionano quelli caratterizzati da $p_t > 20$ GeV e $|n|$ <5. Infine, col paramentro R, ovvero il raggio nel piano $\eta - \phi$ classifichiamo i jet: con raggio R = 0.4 sono definiti AK4 e con raggio R = 0.8 sono definiti AK8.

- **b-tagging**: il quark b, prodotto del decadimento del quark top, adronizza generando un jet. La sua identificazione è fatta tramite algoritmi che sfruttano l'elevata vita media τ dei mesoni B che sono presenti nei jet provenienti dai quark b, che rende loro possbile percorrere una distanza cτ dell'ordine di 450 μm e, quindi, distinguerli dagli altri jet. [18]

- **Elettroni**: gli elettroni vengono individuati nell'ECAL e tracker e vengono selezionati utilizzando una tecnica di identificazione multivariata, in particolare, un decision tree potenziato. [19]

- **Muoni**: i muoni vengono rivelati dalle camere per i muoni e la loro traiettoria è ricostruita anche facendo uso delle informazioni che provengono dal Tracker. Essi vengono classificati in base alla traccia che lasciano:

  - Standalone-muon viene rilevato solo dal sistema a muoni e la sua traiettoria viene ricostruita mediante l'uso dell tecnica del filtro di Kalman.
  - Tracker muon viene rilevato dal silicon tracker ma con tracce nel sistema a muoni. Essi hanno $p_t > 0.5$ GeV e $p > 2.5$ GeV.
  - Global muon viene chiamato cosi il muone le cui le cui tracce del sistema a muoni sono compatibili con quelle del tracker.

  Essi vengono poi distinti in tight e loose a seconda delle loro caratteristiche cinematiche e di richieste imposte sulla qualità della ricostruzione, come una soglia sul χ^2 del fit alla traccia effettuato: [9]

  - Loose muon è un muone selezionato dall'algoritmo particle flow e può essere tracker o un global muon. Sono caratterizzati da $|n| < 2.5$ e $p_t > 10$ GeV.
  - Tight muon può essere solo un global muon a cui gli viene imposto un $\frac{\chi^2}{dof} < 10$. Esso ha $|n| < 2.1$ e $p_t > 26$ GeV.

4.2 Ricostruzione del quark top

Il primo step per la ricostruzione del quark top consiste nella ricostruzione del bosone W. [20] Tutti i suoi prodotti di decadimenti vengono ricostruti dal rivelaori tranne il neutrino che sfugge inosservato, quindi va ricostruito indirettamente. Per la ricostruzione di quest'ultimo, assumiamo che le componenti x e y dell'energia trasversale mancante siano interamente dovute al neutrino in fuga e lo si calcola, imponendo come vincolo la massa di W nella conservazione del'energia totale nel centro di massa.

Il secondo step è la valutazione del 4-momento del quark t. Cruciale è la scelta del AK4 jet, ovvero i jet dei quark b con parametro del cono con R=0.4, che devono essere accoppiati al bosone W e che formano, alla fine, il quark top.

Il CMS ricostruisce quark top in più modi, in base agli oggetti fisici utilizzati per la ricostruzione; in questo lavoro di tesi consideriamo come dati un campione di 10^4 quark t, ricostruiti sia con il neutrino che senza, provenineti da processi di produzione di bosoni W' con una massa di 3 TeV. Nei seguenti grafici della distribuzione della massa e del momento trasverso vengono riportati i valori di segnale e di fondo:

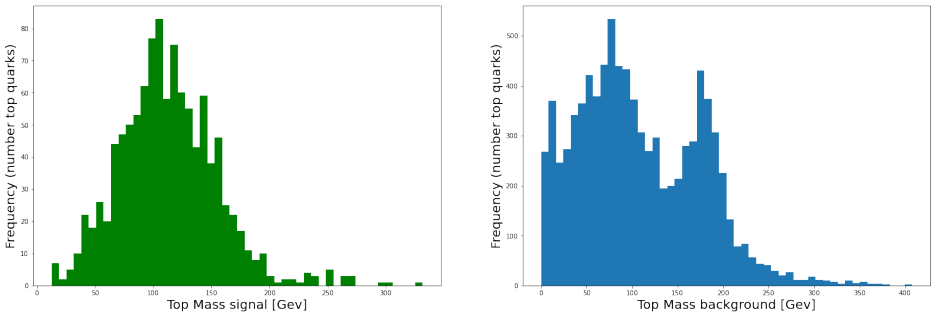


Figura 17: Massa della coppia jet-leptone all'interno di CMS nel caso corrispondente al top vero (sinistra) o falso (destra).

Coi i grafici delle masse si evince come i picchi si avvicinano al valore della massa del quark t prevista dalla teoria, (173.1 ± 1.3) GeV [2], come ci aspettavamo, soprattutto nel caso tel quark t ricostruito col neutrino.

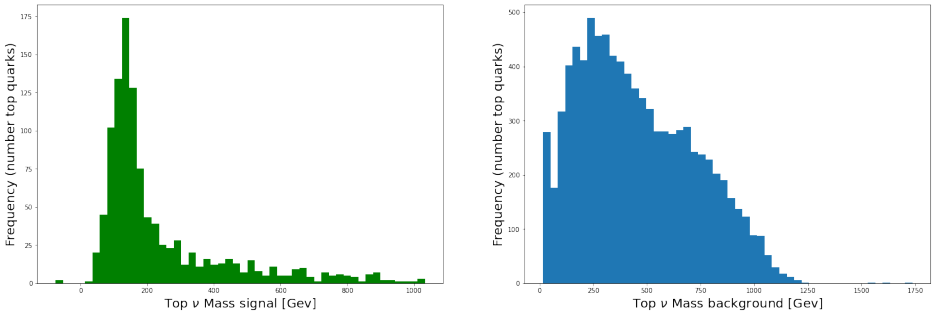


Figura 18: Massa del quark top ricostruito con il neutrino all'interno di CMS nel caso corrispondente al top vero (sinistra) o falso (destra).

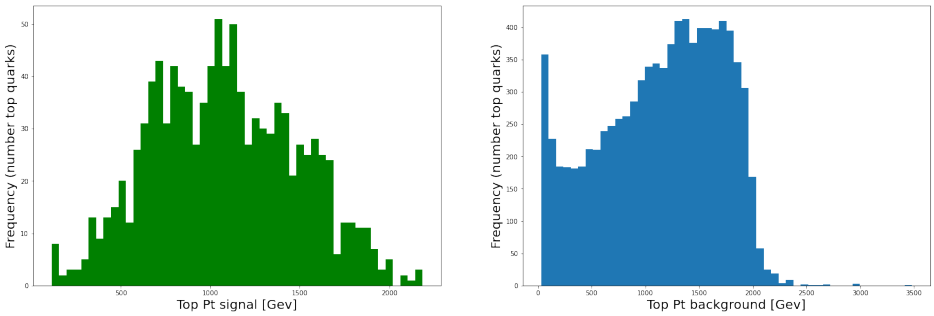


Figura 19: Momento trasverso della coppia jet-leptone all'interno di CMS nel caso corrispondente al top vero (sinistra) o falso (destra).

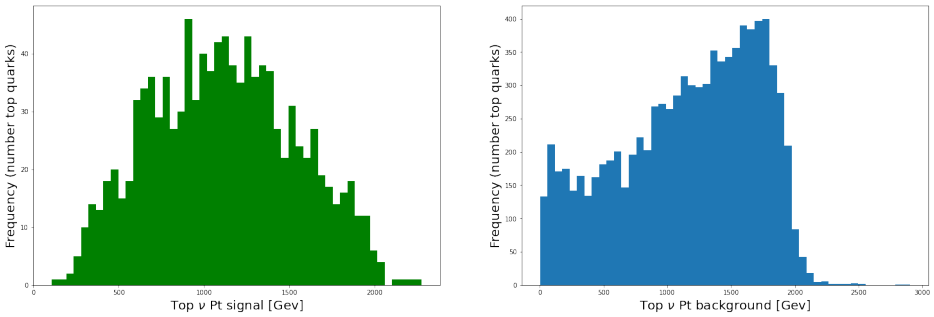


Figura 20: Momento trasverso del quark top ricostruito con il neutrino all'interno di CMS nel caso corrispondente al top vero (sinistra) o falso (destra).

4.3 Analisi delle feature del quark top

Non tutte le feature sono utili allo stesso modo per costruire il quark top. L'identificazione della Feature Importance è stata effettuata impiegndo due algoritmi: il Random Forest e il Boruta. Le feature del quark top prese in esame:

- Jet_{Mass}, Jet_{Pt}, Jet_{Eta}- feature legate alla cinematica dei jet;
- Top_{Mass}, Top_{Pt}, Top_{Eta}- parametri cinematici dei quark top ricostruiti utilizzando la somma dei quadrimoneti del jet e del muone;
- $Muon_{Charge}$, $Muon_{Dxy}$, $Muon_{DxyErr}$, $Muon_{Dz}$, $Muon_{DzErr}$, $Muon_{Pt}$, $Muon_{Eta}$ $Muon_{Phi}$- caratteristiche dei muoni rivelati, in particolare posizione lungo z e nel piano xy del parametro d'impatto della traccia con il relativo errore e quantità cinematiche del muone;
- MET, MET_{phi}- valori dell'energia trasevrsa mancante;
- $Top_{\nu,Mass}$, $Top_{\nu,Pt}$, $Top_{\nu,Eta}$, $Top_{nu,Phi}$- parametri cinematici dei quark top ricostruiti tenendo conto del 4-momento di muoni, jet e neutrini.

4.3.1 Applicazione del Random Forest

Introduciamo un modello di random forest con le seguenti proprietà:

- la massima profondità di ogni albero impostata a 9 nodi;
- il numero di feature da considerare quando si cerca la suddivisione migliore impostata a 2;
- il campione è composto da minimo 20 elementi per ogni nodo foglia;
- i campione che dividono un nodo interno sono minimo 30;
- numero di foglie impostato a 200;
- non controlla né la casualità del bootstrap del campione e né la verbosità quando fitta o predice.

L'allenamento di ogni decision tree è stato effettuato fornendo all'algoritmo il dataset contenente tutti i valori associati ad ogni feature, e i valori di target, cioè il tipo di quark top preso in esame. Dopo allenato, usiamo le feaure importance del random forest; quest'ultimo assegna un valore, ad ogni feature, compreso da 0 a 1, in modo tale che la somma di ogni valore di ogni feature

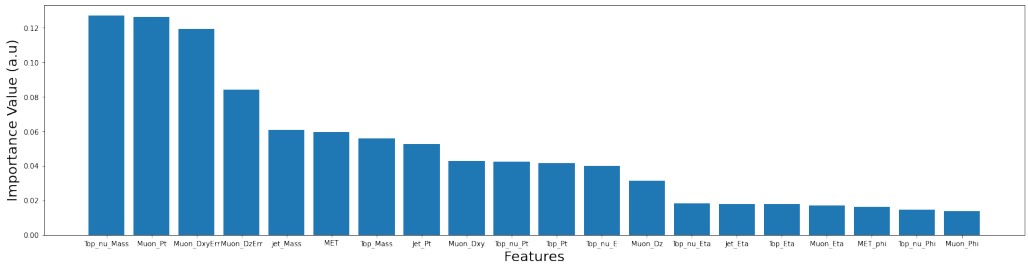


Figura 21: Feature importance del quark top classificate mediante l'analisi con Random Forest

sia 1. Più il valore è prossimo ad 1, più la feature sarà importante. Di seguito riportiamo un grafico delle prime 20 feature più importanti del quark: A questo punto valutiamo l'algoritmo: analizziamo l'accuratezza delle feature importance man mano che si aggiungano le feature più importanti: Si può

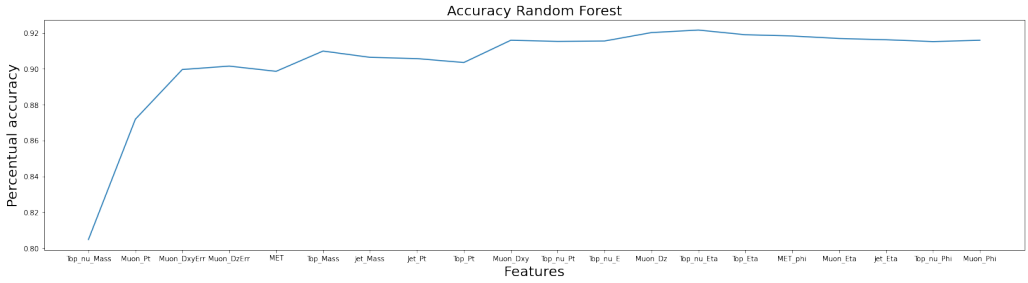


Figura 22: Accuracy della classificazione RF, ottenuta selezionando le feature con importanza crescente secondo il RF stesso

notare che cresce per ogni feature aggiunta fino all'ottava per poi diminuire leggermente, segno che l'allenamento, ormai, non è in grado più di apprendere nulla dato il campione disponibile, ma le perfomance ora fluttuano statisticamente. Inoltre, si nota che le feature vengono classificate e raggruppate non in base alla tipologia (cinematica o di ricostruzione), poichè l'algoritmo procede secondo i citeri di ottimizzazione del random forest. Invece, per la recall e la precision, andiamo a graficare sia per il segnale che per il fondo:

4.3.2 Applicazione di Boruta

L'algoritmo di boruta viene implementato sulla base del random forest utilizzato in precedenza, ed ha le seguenti caratteristiche:

- una soglia per il confronto tra shadow feature e feature reali posta ad un percentile del 90%;

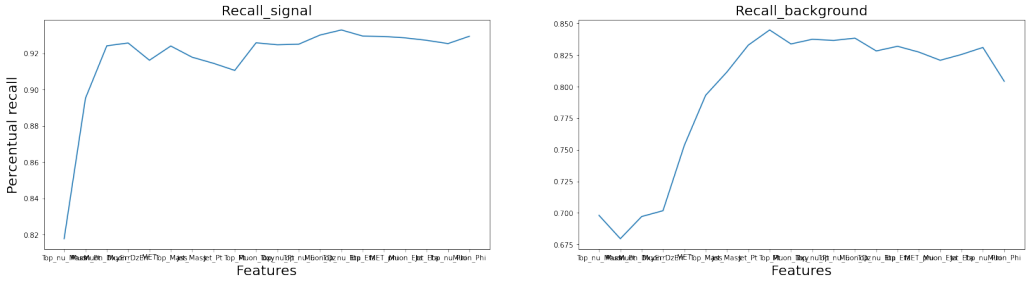


Figura 23: Recall della classificazione RF, ottenuta selezionando le feature con importanza crescente secondo il RF stesso

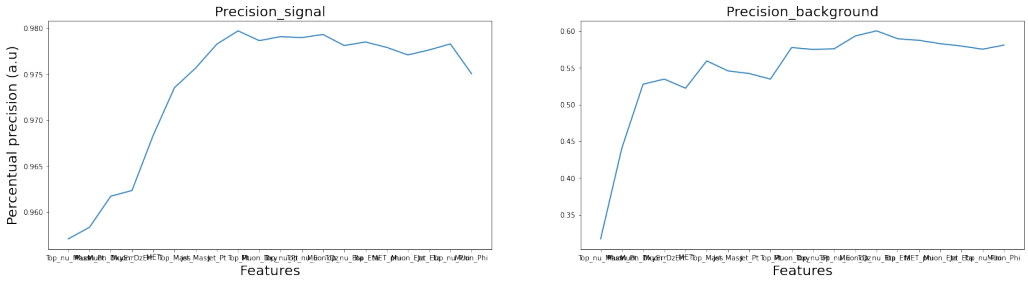


Figura 24: Precision della classificazione RF, ottenuta selezionando le feature con importanza crescente secondo il RF stesso

- un numero stimatori che viene determinato dallo stesso algoritmo in base alla grandezza dei dati forniti;
- il numero massimo di iterazioni impostato a 100.

Con boruta le feature, dopo essere state valutate, vengono etichettate con un valore, che chiameremo rango, che va da 1 (valore che corrisponde alla feature più importante) a seguire. Ogni valore di rango corrisponde ad una soglia superata da parte della feature corrispondente. Un rango basso indica una soglia di accettazione alta, che poi diminuisce man che mano che l'algoritmo individua variabili meno importanti. Le feature selezionte con boruta

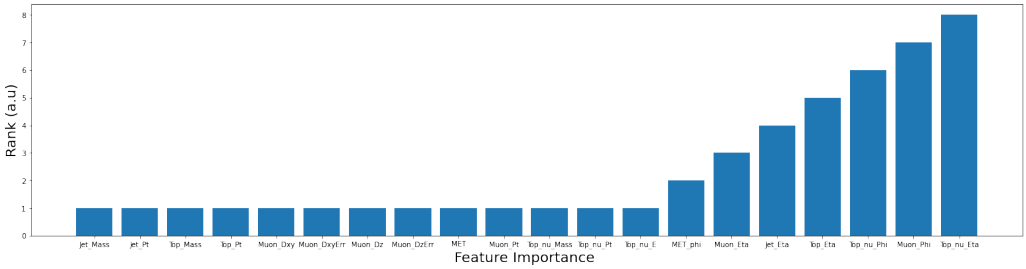


Figura 25: Feature importance del quark top, classificate mediante la variabile rango ottenuta tramite analisi con Boruta

vengono analizzate e raggruppate dall'algoritmo non in maniera casuale ma con un criterio: si individuano dei gruppi di feature caratteristiche. Il primo gruppo, da sinistra a destra, ha in sè le feature cinematiche dei jet (Jet_{Mass} e Jet_{Pt}), per poi passare al gruppo che contiene parametri legati alla costruzione dei top (Top_{mass} e Top_{pt}), a quelli legati ai muoni (da $Muon_{Dxy}$ a $Muon_{Pt}$) e, per finere, con le caratteristiche dei top ricostruiti con il neutrino (da $Top_{\nu,mass}$ a $Top_{\nu,E}$). Ciò risulta evidente nei grafici di accuracy, recall e precision in funzione delle varibili individuate.

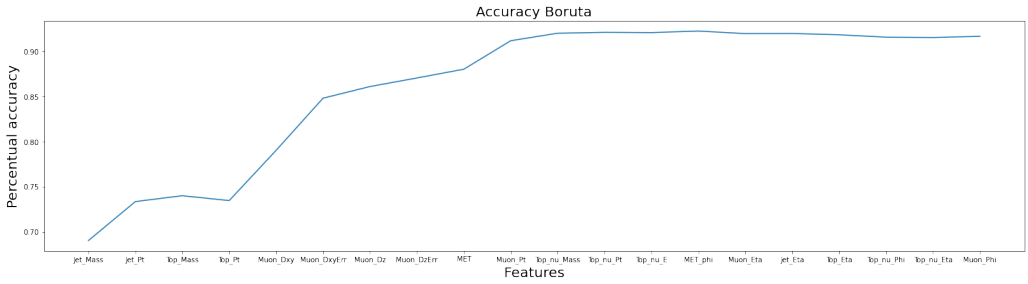


Figura 26: Accuracy della classificazione RF ottenuta selezioanando le feature con Boruta in ordine crescente di importanza.

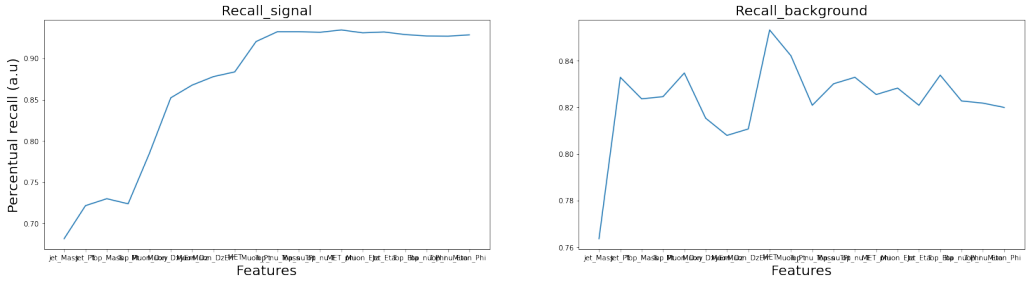


Figura 27: Recall della classificazione RF ottenuta selezioanando le feature con Boruta in ordine crescente di importanza.

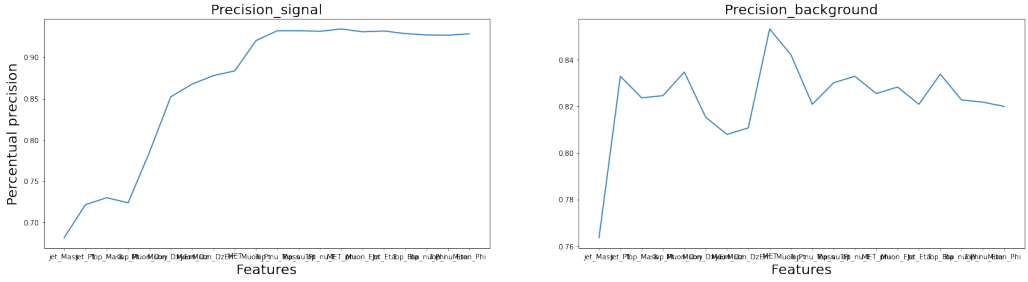


Figura 28: Precision della classificazione RF ottenuta selezioanando le feature con Boruta in ordine crescente di importanza.

4.4 Ottimizzazione del quark top

Con gli stessi due modelli utilizzati in precedenza, identifichiamo i quark top, in questo caso con la classificazione delle feature importance fornita da entrambi i modelli e, con questi quark top, ricostruiamo il bosone W'. Per il RF prediamo le prime 20 variabili, per Boruta quelle con rango uguale a 1.

Consideriamo 10^4 eventi simulati, in cui si hanno la produzione di diversi oggetti fisici, tra cui i quark top ottenuti mediante la somma dei 4-momenti dei jet e dei muoni dell'evento. Per ogni evento, in cui si formano vari quark top, selezioniamo quelli che hanno il valore dello score più alto fornito dai modelli, applicando al valore dell'output del discriminatore una stessa soglia, scelta in maniera larga in corrispondeza ad una probabilità del 25% che il quark t fosse corretto, secondo l'algoritmo. Di questi quark ricaviamo la massa e il momento trasverso e li riportiamo nei seguenti grafici:

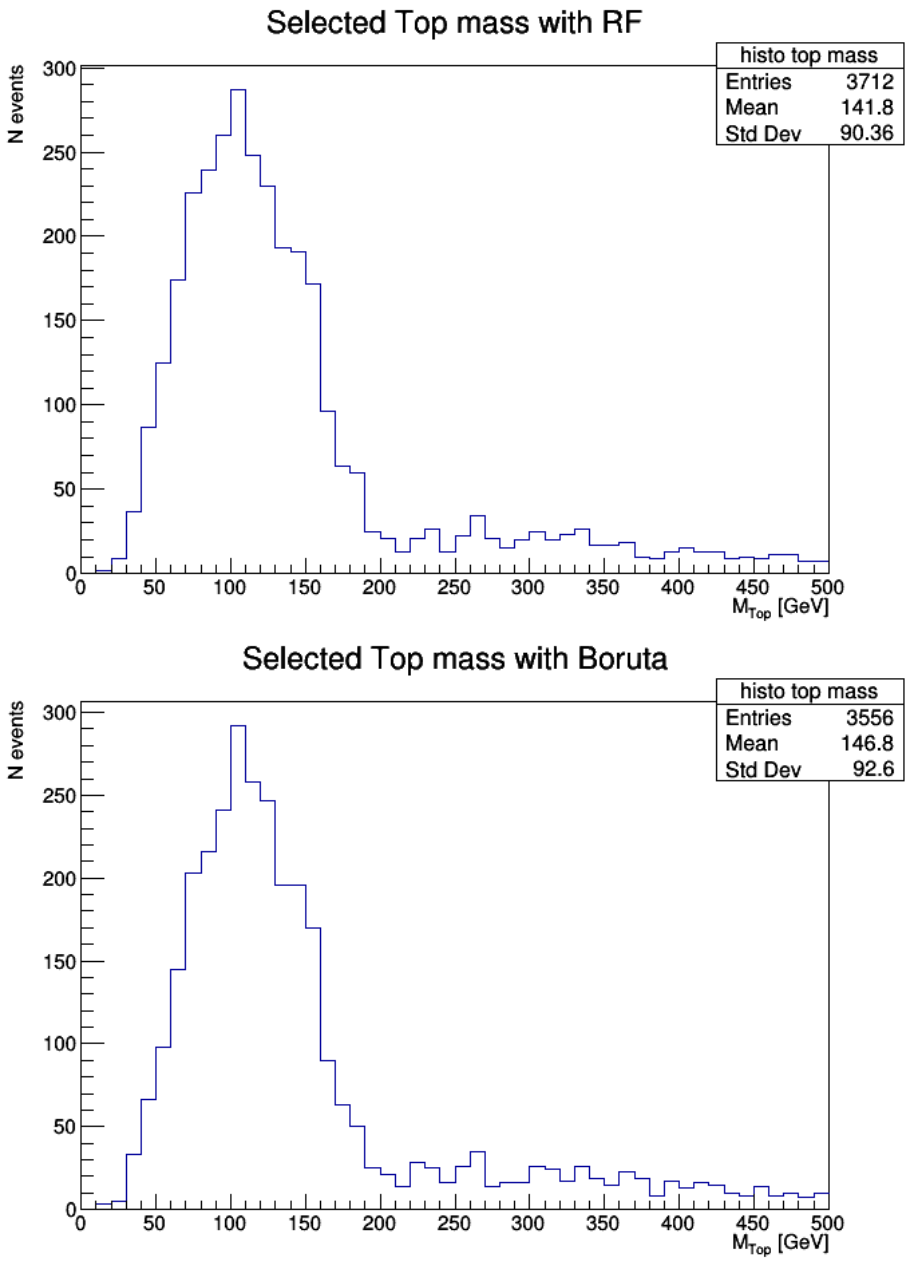


Figura 29: Massa del quark top, ricostruito come coppia jet-muone, con le feature importance trovate dal Random Forest (in alto) e Boruta (in basso).

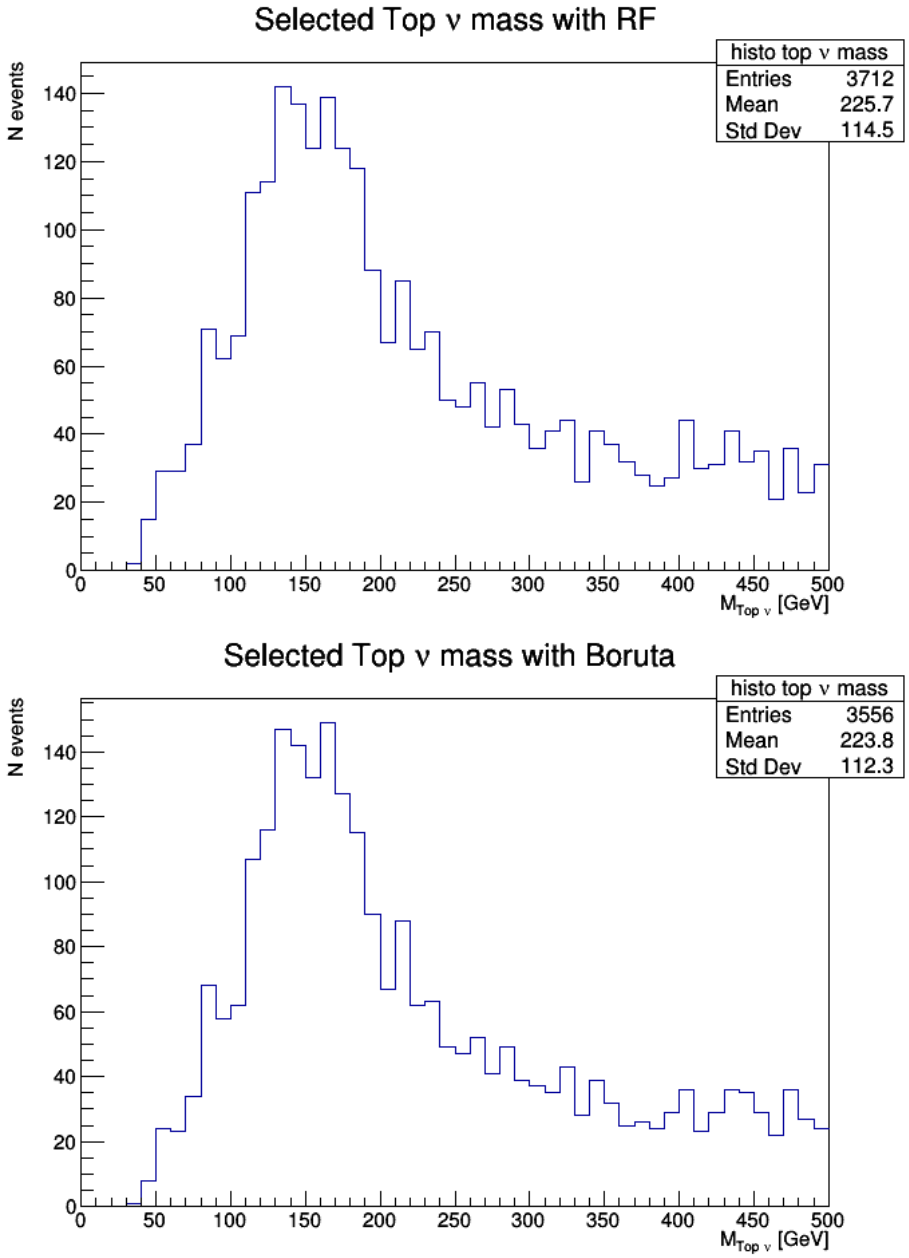


Figura 30: Massa del quark top, ricostruito includendo il neutrino, con le feature importance trovate dal Random Forest (in alto) e Boruta (in basso).

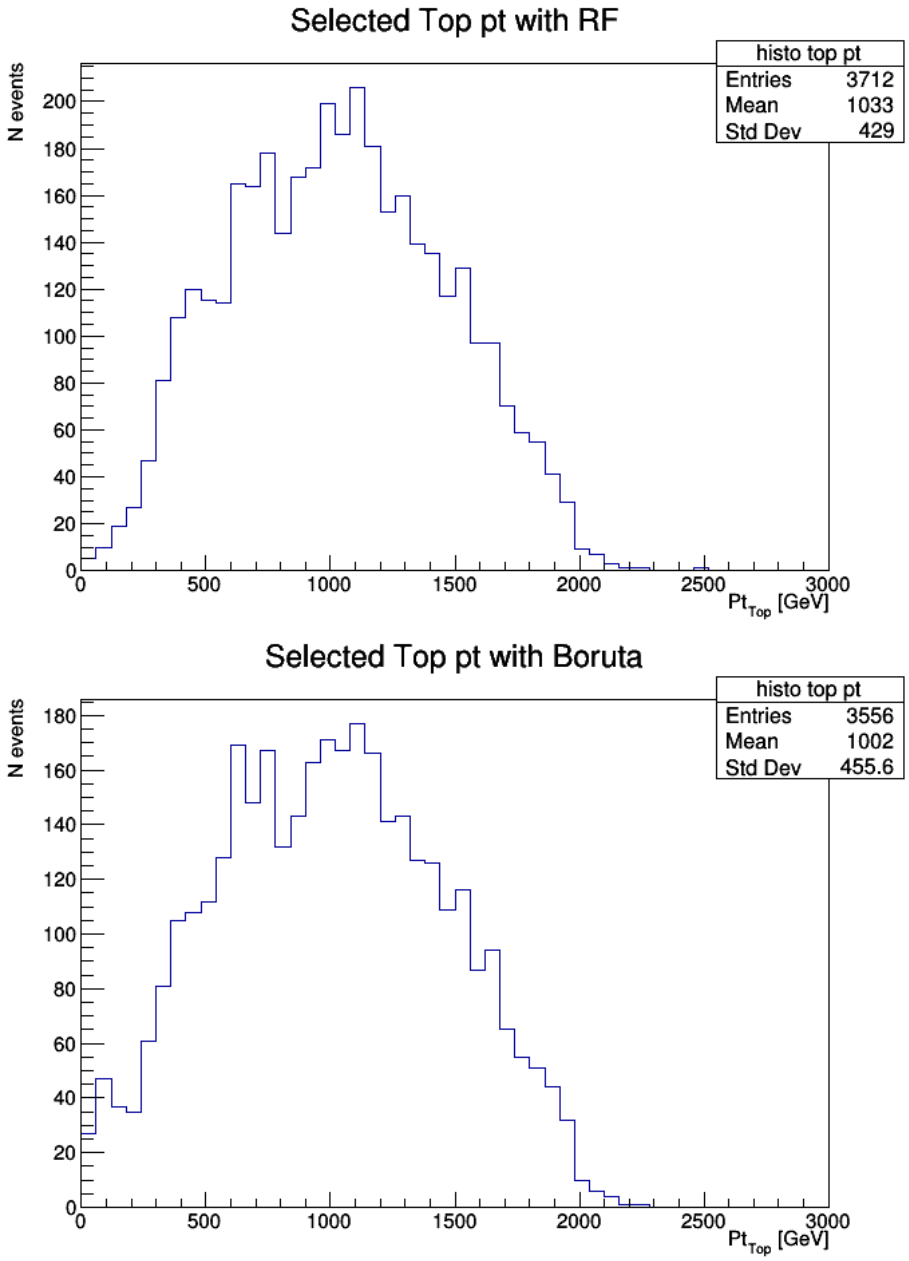


Figura 31: Momento trasverso del quark top, ricostruito come coppi jet-muone, con le feature importance trovate dal Random Forest (in alto) e Boruta (in basso).

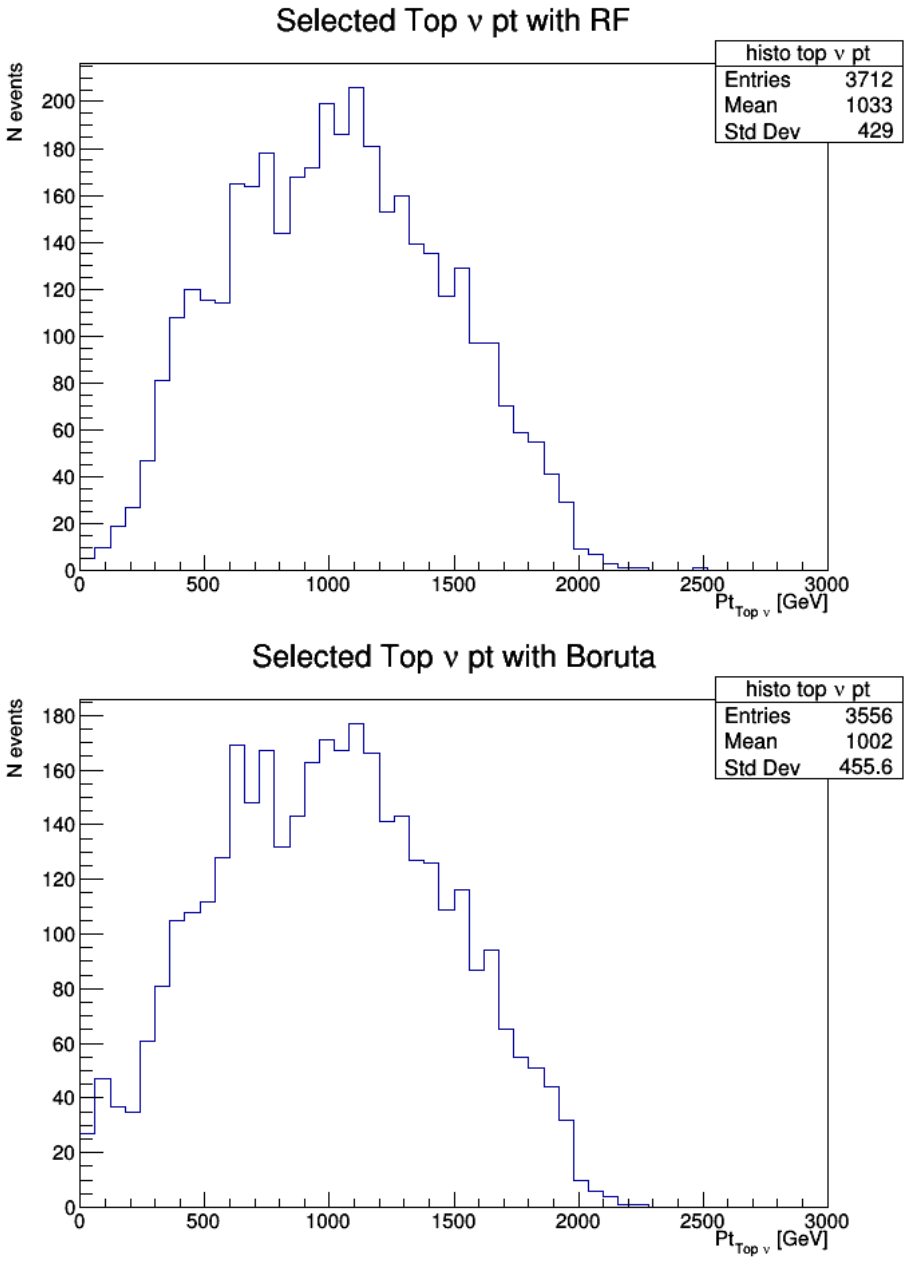


Figura 32: Momento trasverso del quark top, ricostruito includendo il neutrino, con le feature importance trovate dal Random Forest (in alto) e Boruta (in basso).

4.5 Ricostruzione del bosone W'

Per la ricostruzione del bosone W', abbiamo considerato, oltre al top, il jet proveniente dal quark b prodotto nel decadimento del suddetto bosone. Si è selezionato il jet più energetico con la condizione non facesse parte del quark top selezionato in precedenza. La ricostruzione della massa è stata fatta tramite la conservazione del quadrimomento e si può notare come i picchi dei due grafici di seguito siano in prossimità dei 3 TeV, come ci si aspettava.

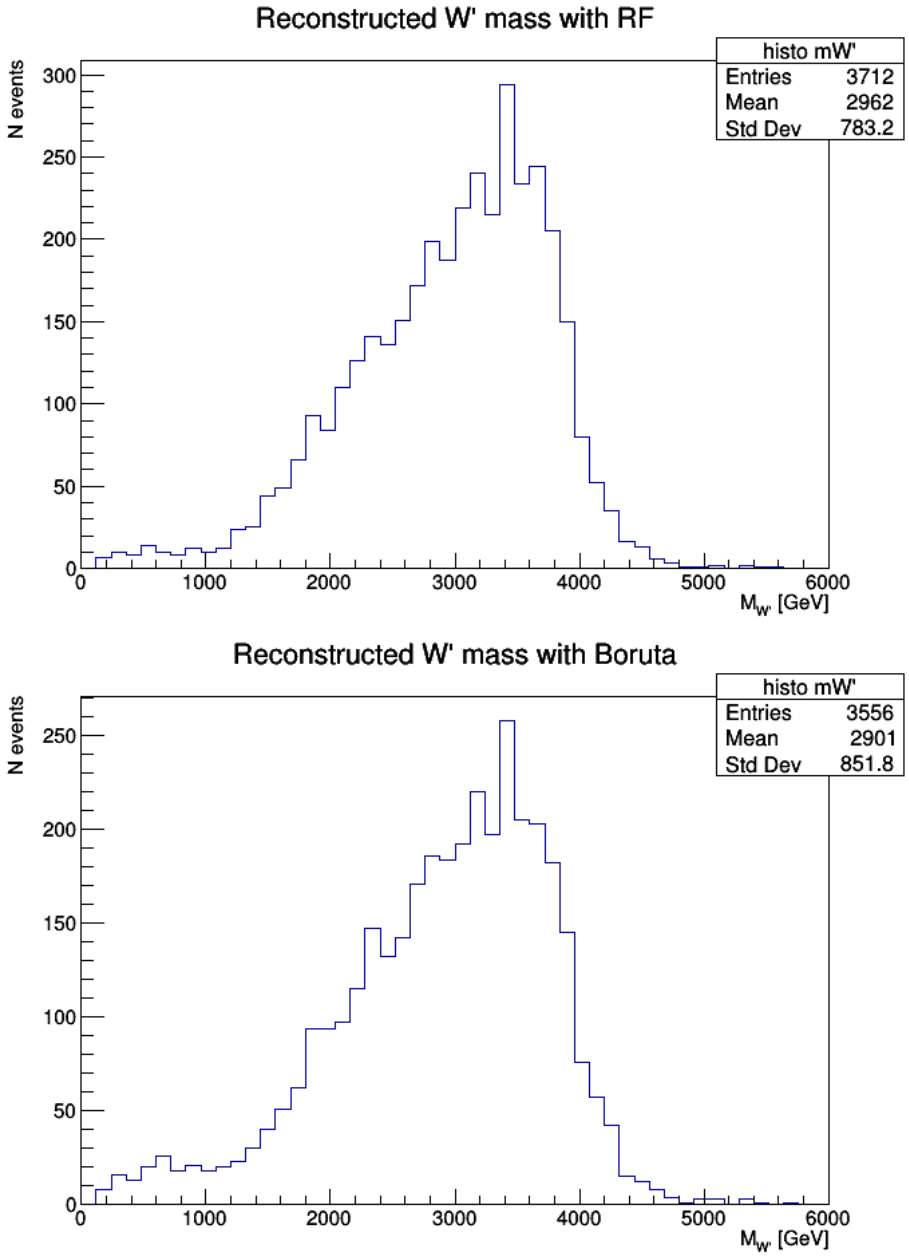


Figura 33: Massa del bosone W', ricostruito con il quark top ricavato con le feature importance trovate dal Random Forest (in alto) e Boruta (in basso).

Conclusione

In questo lavoro di tesi è stato effettuato uno studio delle feature del quark top, a seguito del decadimento di un nuovo ipotetico bosone carico, chiamato W', oggetto delle ricerche fatte a LHC con l'esperimento CMS. Questo nuovo tipo di bosone è ipotizzato da molte teorie fisiche che si propongono di estendere il Modello Standard, come il left-right symmetric model e i modelli little Higgs, con l'intento di superare il MS che mostra alcune criticità.

I processi di produzione del bosone W' e le rispettive segnature nel rivelatore CMS sono stati simulati con metodi di simulazione Monte Carlo. Il quark t non può essere direttamente individuato, ma le sue caratteristiche possono essere ricavate dai prodotti finali della catena di decadimento: $t \to Wb\ W \to \mu\nu_\mu$, all'interno del CMS. L'analisi standard, effettuata da CMS, ricava il 4-momento del top sommando i 4-momenti dei prodotti finali e in questo modo ottiene le proprietà cinematiche del quark t.

Per migliorare la classificazione delle feature del quark top sono stati considerati due algoritmi di ML: il random forest e boruta, per classificare le feature importance del quark t, il che ha permesso di apprezzare le differenze nei due algoritmi.

Infine, ottenuta la feature importance, abbiamo ricavato le proprietà cinematiche del quark t, selezionando, nei vari eventi simulati, solo candidati quark t che superavano una selezione basata sugli algoritmi allenati con tali feature. Sommando i 4-momenti dei quark t ricavati in questo modo con il b-jet più energetico, abbiamo, in definita, ricavato la distribuzione della massa del bosone W', riproducendo fedelmente la distribuzione attesa.

Un possibile lavoro di miglioramento potrebbe consistere nello studio più dettagliato delle soglie di selezione dei quark top con l'aggiunta del canale elettronico. Un'ulteriore rifinitura sarebbe lo studiare altri algoritmi di machine learning con una performance maggiore da applicare ai futuri dati di collisione di LHC

Riferimenti bibliografici

[1] Paul Langacker. Introduction to the standard model and electroweak physics. *arXiv preprint arXiv:0901.0241*, 2009.

[2] P.A. Zyla et al. Review of Particle Physics. *PTEP*, 2020(8):083C01, 2020.

[3] María Josefina Alconada Verzini, Francisco Alonso, Francisco Anuar Arduh, María Teresa Dova, Joaquín Hoya, Fernando Gabriel Monticelli, Hernán Pablo Wahlberg, ATLAS Collaboration, et al. Search for w'→ tb decays in the hadronic final state using pp collisions with energy in the center of mass at 13 tev with the atlas detector. *Physics Letters B*, 781, 2018.

[4] CMS collaboration et al. Searches for w'bosons decaying to a top quark and a bottom quark in proton-proton collisions at 13 tev. *arXiv preprint arXiv:1706.04260*, 2017.

[5] Lucio Rossi. Nuova luce per nuova fisica: al cern posa della prima pietra di hilumi lhc, 2018. [Online; accessed 3-1-2021].

[6] Lyndon Evans, Philip Bryant, and LHC Machine. Jinst 3. *S08001*, pages 1748–0221, 2008.

[7] CMS Collaboration Collaboration et al. The cms electromagnetic calorimeter project: Technical design report. *Technical Design Report CMS. CERN, Geneva*, 1997.

[8] Pierluigi Paolucci. L'esperimento cms ad lhc, 2011. [Online; accessed 7-01-2022].

[9] Albert M Sirunyan, CMS Collaboration, et al. Performance of the cms muon detector and muon reconstruction with proton-proton collisions at root s= 13 tev. 2018.

[10] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[11] Josep Lluís Solé. Book review: Pattern recognition and machine learning. cristopher m. bishop. information science and statistics. springer 2006, 738 pages., 2007.

[12] Andrea De Mauro. *Big Data Analytics: Analizzare e interpretare dati con il machine learning*. Apogeo editore, 2019.

[13] J. Ross Quinlan. Simplifying decision trees. *International journal of man-machine studies*, 27(3):221–234, 1987.

[14] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning*. Springer series in statistics New York, 2001.

[15] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[16] Aruna Nayak, Pedrame Bargassa, Cristóvão Beirão Da Cruz E Silva, Agostino Di Francesco, Pietro Faccioli, Bruno Galinhas, Michele Gallinaro, Jonathan Hollar, Nuno Leonardo, Lara Lloret Iglesias, et al. Search for heavy resonances decaying to a top quark and a bottom quark in the lepton+ jets final state in proton–proton collisions at 13 tev. 2018.

[17] Albert M Sirunyan, Armen Tumasyan, Wolfgang Adam, Ece Asilar, Thomas Bergauer, Johannes Brandstetter, Erica Brondolin, Marko Dragicevic, Janos Erö, Martin Flechl, et al. Particle-flow reconstruction and global event description with the cms detector. *Journal of Instrumentation*, page 86, 2017.

[18] Albert M Sirunyan, Malte Backhaus, Lukas Bäni, Pirmin Berger, Lorenzo Bianchini, Günther Dissertori, Michael Dittmar, Mauro Donegà, Christian Dorfer, Christoph Grab, et al. Identification of heavy-flavour jets with the cms detector in pp collisions at 13 tev. *Journal of Instrumentation*, 13:P05011, 2018.

[19] CMS collaboration et al. Performance of electron reconstruction and selection with the cms detector in proton-proton collisions at sqrt (s)= 8 tev. *arXiv preprint arXiv:1502.02701*, 2015.

[20] CMS collaboration et al. Search for w'to tb decays in the lepton+ jets final state in pp collisions at sqrt (s)= 8 tev. *arXiv preprint arXiv:1402.2176*, 2014.

Ringraziamenti

Le persone da ringraziare per questi tre anni di università sono veramente tante, perchè merita un piccolo ringraziamento anche chi mi ha passato una sola pagina di appunti che non avevo, chi mi ha passato solo un esercizio che non ero riuscito a svolgere o chi mi ha fatto sorridere anche per un secondo. Quindi, per non citare cento persone, e poichè già ho sforato le 40 pagine massime di tesi, sarò molto breve: ringrazio i miei due relatori: il professor Luca Lista e il dottor Orso Alberto Maria Iorio per l'infinita pazienza e cordialità e il collega Antonio Marzullo per aver vissuto, praticamente, insieme questa avventura.