



UNIVERSIDADE FEDERAL DA BAHIA  
ESCOLA POLITÉCNICA  
DEPARTAMENTO DE ENGENHARIA QUÍMICA

AMANDA SANTIAGO ACÁCIO

**CALIBRAÇÃO DAS MEDIÇÕES DE ENERGIA DE UM DETECTOR DE PARTÍCULAS  
UTILIZANDO ÁRVORE DE DECISÃO COM IMPULSO POR GRADIENTE**

SALVADOR – BA



AMANDA SANTIAGO ACÁCIO

**CALIBRAÇÃO DA MEDIÇÃO DE ENERGIA DE UM DETECTOR DE PARTÍCULAS  
UTILIZANDO ÁRVORE DE DECISÃO COM IMPULSO POR GRADIENTE**

Trabalho de Conclusão de Curso apresentado ao Curso de graduação em Engenharia de Controle e Automação de Processos, Escola Politécnica, Universidade Federal da Bahia, como requisito parcial para obtenção do grau de Bacharel em Engenharia de Controle e Automação.

Orientador: Professor Dr. Eduardo F. de Simas Filho

SALVADOR – BA  
2023

AMANDA SANTIAGO ACÁCIO

**CALIBRAÇÃO DA MEDIÇÃO DE ENERGIA DE UM DETECTOR DE PARTÍCULAS  
UTILIZANDO ÁRVORE DE DECISÃO COM IMPULSO POR GRADIENTE**

Trabalho de conclusão de curso apresentado ao Colegiado do curso de Graduação em Engenharia de Automação e Controle de Processos da Universidade Federal da Bahia como requisito para a obtenção do título de Bacharel em Engenharia de Automação e Controle de Processos.

14 de Julho de 2023.

Comissão examinadora:

---

Prof. Dr. Eduardo F. de Simas Filho  
Universidade Federal da Bahia

---

Me Paulo Roberto A. da Silva  
Petrobrás e Doutorando no PPGEE/UFBA

---

Me Bassem Yousseff Makhoul Jr  
Doutorando no PPGEE/UFBA

SALVADOR - BA

2023

## **AGRADECIMENTOS**

Agradeço primeiramente a Deus pela sabedoria, força e paciência durante todo o processo de graduação.

Também expesso minha imensa gratidão a minha mãe Marinice e ao meu pai Ubiratan, pelo suporte emocional e encorajamento ao longo de toda a jornada acadêmica. O apoio e as palavras de encorajamento que recebi de vocês foram um combustível indispensável para seguir em frente.

Agradeço a minha família por todo o apoio e incentivo durante todo esse período, e a todos os meus amigos pela compreensão, paciência, acolhimento e ajuda. Sem vocês essa jornada teria sido muito mais difícil.

Agradeço ao meu orientador Eduardo Simas, cuja orientação cuidadosa, e apoio constante foram fundamentais para a conclusão deste estudo.

Por fim, sou grata a todas as pessoas que, de alguma forma, contribuíram para a conclusão desse trabalho, direta ou indiretamente.

*Nossas virtudes e nossos defeitos são  
inseparáveis, assim como a força e a matéria.  
Quando se separam, o homem deixa de existir.*

Nikola Tesla, 1902. Engineer.

## RESUMO

Este trabalho apresenta a análise de um modelo de calibração utilizando árvores de decisão na etapa rápida do *High-Level Trigger* (HLT) do Experimento ATLAS, que é um componente essencial do sistema de filtragem de dados e tomada de decisão do no Grande Colisor de Hádrons (do inglês: *Large Hadron Collider*). No Experimento ATLAS, o processo de calibração das medições de energia no HLT costuma ser feito com GBDT (*Gradient Boosted Decision Tree*). Diante disso, o objetivo desse estudo é avaliar o desempenho e comportamento da GBDT para diferentes conjuntos de entradas. Na primeira parte do trabalho, é apresentada uma revisão teórica sobre os princípios de funcionamento do LHC e do Experimento ATLAS, bem como a sua importância no contexto da física de partículas. São abordados os aspectos relacionados à necessidade de filtragem online e o papel do HLT na seleção de eventos de interesse científico. A segunda parte do trabalho se concentra nos aspectos técnicos e operacionais do HLT do Experimento ATLAS. São apresentados os principais métodos e técnicas utilizados, como o uso de árvores de decisão para a classificação dos eventos de interesse. Além disso, são discutidas as estratégias de otimização na implementação e no ajuste da GBDT no HLT. Por fim, são apresentados resultados e análises de desempenho da GBDT, incluindo análise gráfica da dispersão, e variância dos treinamentos realizados. Essas análises são fundamentais para avaliar a capacidade da GBDT em cumprir os objetivos de otimizar as demandas computacionais e melhorar a eficiência na seleção de partículas eletromagnéticas. Os resultados mostraram que a diferença relativa entre a configuração de entrada da GBDT com vinte e oito anéis e a configuração de cem anéis não ultrapassaram 10%, demonstrando que a estimativa dos eventos para os dois arranjos são similares.

**Palavras-chave:** Experimento ATLAS, sistema de filtragem, calibração, HLT, GBDT

## ABSTRACT

This work presents the analysis of a calibration model using decision trees in the High-Level Trigger (HLT) fast stage of the ATLAS experiment, which is an essential component of the data filtering and decision making system of the Large Hadron Collider. In the ATLAS Experiment, the calibration process of energy measurements in the HLT is usually done with GBDT (Gradient Boosted Decision Tree). Given this, the objective of this study is to evaluate the performance and behavior of GBDT for different sets of inputs. In the first part of the paper, a theoretical review is presented on the principles of operation of the LHC and the ATLAS Experiment, as well as their importance in the context of particle physics. It addresses aspects related to the need for filtering and the role of the HLT in the selection of scientifically interesting events. The second part of the work focuses on the technical and operational aspects of the HLT of the ATLAS detector. The main methods and techniques used, such as artificial neural networks and decision trees for event classification, are presented. In addition, optimization strategies in the implementation and tuning of the GBDT in the HLT are discussed. Finally, results and performance analyses of the GBDT are presented, including graphical analyses of dispersion and variance of the conducted trainings. These analyses are crucial to evaluate the capability of the GBDT in achieving the objectives of optimizing computational demands and improving the efficiency in the selection of electromagnetic particles. The results showed that the relative difference between the twenty-eight-ring GBDT input configuration and the one hundred-ring configuration did not exceed 10%, demonstrating that the estimation of events for the two arrangements are similar.

**Keywords:** ATLAS detector, filtering system, calibration, HLT, GBDT.

## LISTA DE FIGURAS

Figura 1.1 – Túnel do acelerador de prótons e antiprótons do CERN.....	2
Figura 1.2 – Representação da disposição dos detectores de partículas ao longo de <i>Lareg Hardron Collider</i> .....	3
Figura 2.1 – Vista em corte do Experimento ATLAS e seus respectivos subdetectores.	9
Figura 2.2 – Diagrama ilustrativo do sistema de calorimetria do Experimento ATLAS..	10
Figura 2.3 – Diagrama de blocos que representa as etapas que ocorrem no processo de filtragem do sistema de trigger L1 do Experimento ATLAS.....	13
Figura 2.4 – Diagrama que representa o fluxo de dados no segundo nível de filtragem online do Experimento ATLAS. ....	16
Figura 2.5 – Diagrama de blocos que representa a geração de dados do ATLAS para análise via ATHENA.....	17
Figura 2.6 – Representação do sistema de coordenadas do Experimento ATLAS.....	19
Figura 2.7 – Diagrama do sistema de coordenadas cilíndricas do ATLAS.....	19
Figura 2.8 – Diagrama de um corte transversal do detector que representas as suas camadas e as possíveis trajetórias de um evento. ....	22
Figura 2.9 – Representação dos anéis e suas diferentes granularidades de quatro camadas do calorímetro. Os pontos em vermelho, são as células mais energéticas. ..	23
Figura 2.10 – Diagrama de um algoritmo de aprendizado de máquina. Nesse caso, são fornecidos N recursos de entrada, e o método multivariado (MVA) atribui cada evento da amostra a uma das três classes.....	24
Figura 2.11 – Representação de um conjunto de dados e a sua árvore de decisão correspondente. As elipses ou bolhas são os nós de decisão e os retângulos são os nós de folha. O nó de decisão se subdivide até que as classes estejam separadas.....	27
Figura 3.1 – Diagrama que ilustra a integração do bloco de calibração proposto com o HLT (High-Level Trigger) do ATLAS.....	31
Figura 3.2 – Diagrama que representa a etapa do sistema de calibração incluído no segundo nível de filtragem do Experimento ATLAS. ....	32
Figura 3.3 – Representação das etapas de operação do sistema de calibração proposto. ....	33
Figura 4.1 – Gráfico comparativo da IQR para diferentes intervalos de $\eta$ (pseudorapidez) para a calibração padrão, e para a BDT treinada com 28 e 100 anéis. ....	37

Figura 4.2 – Gráfico comparativo da IQR para diferentes intervalos de $E_t$ para a calibração padrão, e para a BDT treinada com 28 e 100 anéis. ....	38
Figura 4.3 – Histograma do erro relativo das energias estimadas sem a BDT e com a BDT com as configurações de entrada <i>Rings28</i> e <i>Rings100</i> – intervalo de [5,8] GeV.....	39
Figura 4.4 – Histograma do erro relativo das energias estimadas sem a BDT e com a BDT com as configurações de entrada <i>Rings28</i> e <i>Rings100</i> em escala logarítmica – intervalo de [5,8] GeV. ....	40
Figura 4.5 – Histograma do erro relativo das energias estimadas sem a BDT e com a BDT com as configurações de entrada <i>Rings28</i> e <i>Rings100</i> .escrever – intervalo de [20,30] GeV. ....	41
Figura 4.6 – Histograma do erro relativo das energias estimadas sem a BDT e com a BDT com as configurações de entrada <i>Rings28</i> e <i>Rings100</i> em escala logarítmica – intervalo de [20,30] GeV. ....	41
Figura 4.7 – Histograma do erro relativo das energias estimadas sem a BDT e com a BDT com as configurações de entrada <i>Rings28</i> e <i>Rings100</i> – intervalo de [100,150] GeV. ....	42
Figura 4.8 – Histograma do erro relativo das energias estimadas sem a BDT e com a BDT com as configurações de entrada <i>Rings28</i> e <i>Rings100</i> em escala logarítmica – intervalo de [100,150] GeV. ....	43
Figura 4.9 – Histograma do erro relativo das energias estimadas sem a BDT e com a BDT com as configurações de entrada <i>Rings28</i> e <i>Rings100</i> .escrever – intervalo de [150,200] GeV. ....	43
Figura 4.10 – Histograma do erro relativo das energias estimadas sem a BDT e com a BDT com as configurações de entrada <i>Rings28</i> e <i>Rings100</i> em escala logarítmica – intervalo de [150,200] GeV.....	44
Figura 4.11 – <i>Scatter plot</i> da calibração com a BDT com as configurações de entrada <i>Rings28</i> e <i>Rings100</i> – intervalo de [5,8] GeV.....	45
Figura 4.12 – <i>Scatter plot</i> da calibração com a BDT com as configurações de entrada <i>Rings28</i> e <i>Rings100</i> – intervalo de [20,30] GeV.....	46
Figura 4.13 – <i>Scatter plot</i> da calibração com a BDT com as configurações de entrada <i>Rings28</i> e <i>Rings100</i> – intervalo de [100,150] GeV.....	47
Figura 4.14 – <i>Scatter plot</i> da calibração com a BDT com as configurações de entrada <i>Rings28</i> e <i>Rings100</i> – intervalo de [150,200] GeV.....	47
Figura 4.15 – Histograma da diferença relativa das energias estimadas com a BDT com as configurações de entrada <i>Rings28</i> e <i>Rings100</i> . ....	48

## LISTA DE TABELAS

Tabela 2.1 – Número de anéis por camada do sistema de calorimetria do ATLAS. ....	22
Tabela 3.1 – Quantidades de anéis selecionados para cada camada. ....	34

## SUMÁRIO

<b>1. INTRODUÇÃO .....</b>	<b>1</b>
1.1. Motivação .....	5
1.2. Objetivos.....	7
1.2.1. Objetivo Geral .....	7
1.2.2. Objetivos Específicos: .....	7
1.3. Organização do Trabalho .....	8
<b>2. FUNDAMENTAÇÃO TEÓRICA .....</b>	<b>9</b>
2.1. Experimento ATLAS .....	9
2.2. Sistema de Seleção Online .....	11
2.3. Primeiro Nível de Filtragem .....	12
2.4. Segundo Nível de Filtragem .....	14
2.5. Identificação dos Elétrons.....	18
2.6. Neural Ringer .....	21
2.7. Algoritmos de Aprendizado de Máquina .....	23
2.8. Árvores de Decisão .....	25
<b>3. METODOLOGIA.....</b>	<b>29</b>
<b>4. RESULTADOS .....</b>	<b>36</b>
4.1 Análise dos Resultados .....	49
<b>5. CONCLUSÃO.....</b>	<b>50</b>
5.1. Trabalhos Futuros .....	50
<b>REFERÊNCIAS.....</b>	<b>52</b>

## 1. INTRODUÇÃO

A Física de partículas tem como objetivo estudar os componentes elementares da matéria e como eles interagem. Apesar das especulações de Demócrito de que a matéria era constituída de pequenas partículas indestrutíveis, e da teoria atômica desenvolvida por Dalton, que define o átomo como uma partícula esférica, maciça e indivisível, os estudos acerca da física de partículas teve seu início com a descoberta em 1897 do elétron por Thomson, seguida pela descoberta do núcleo atômico por Rutherford (TIPLER; LIEWELLYN, 2017).

À medida que os estudos avançaram, novas informações foram surgindo como a descoberta do nêutron em 1932, do fóton em 1900 por Planck., do efeito fotoelétrico em 1905, do pósitron em 1930, múon em 1937, e várias outras partículas que compõem o universo. Assim, para que esses conceitos físicos sejam validados, são realizados experimentos que necessitam de processamento complexos, velozes e com alta dimensionalidade de eventos (GRIFFITHS, 2020).

Uma consideração importante desses experimentos é a distinção entre partículas elementares e partículas compostas. Historicamente, diversas partículas foram consideradas elementares (que não podem ser divididas em constituintes menores), mas ao colidirem, foram descobertos novos tipos partículas. Essas descobertas levaram ao desenvolvimento do Modelo Padrão, que descreve e prevê as características e interações de todas as partículas com base em um conjunto reduzido de partículas que, até o momento, são consideradas elementares. (THOMSON, 2013).

A descoberta desses novos constituintes da matéria, foram potencializadas, principalmente através do uso dos aceleradores de partículas. Os aceleradores de partículas são equipamentos que aceleram feixes de partículas carregadas, que ao alcançarem altas velocidades e colidirem, possibilitam a detecção e medição das propriedades fundamentais da matéria (MARTINS; SILVA, 2014).

Os aceleradores de partículas podem ser classificados em lineares ou circulares. Nos lineares as partículas carregadas ganham energia deslocando-se em

linha reta, por um campo elétrico estacionário ou através de campos elétricos com radiofrequência. Os circulares, por sua vez, ao invés de acelerar partículas em uma seção reta mantêm o feixe em uma trajetória circular, sendo este acelerado a cada volta (MARTINS; SILVA, 2014).

O feixe de partículas acelerado, após obter a energia desejada pode ser colidido com outro feixe de partículas ou em alvo fixo para que a sua estrutura possa ser estudada. Os aceleradores destinados a isso são os síncrotron e são chamados anéis colisores. O LHC (*Large Hadron Collider*) é o maior acelerador síncrotron em operação atualmente e foi construído com o objetivo de analisar a estrutura fundamental da matéria e investigar as propriedades das partículas fundamentais propostas pelo Modelo Padrão (MOREIRA, 2009; PIMENTA et al., 2013).

O LHC, representado na Figura 1.1 possui 27 km de circunferência que acelera feixe de prótons a 7 TeV e fica enterrado a aproximadamente 100 m de profundidade, entre as cidades de Genebra na Suíça e Meyran na França, sendo o CERN (Organização Europeia para a Pesquisa Nuclear), o responsável pela sua operação (CERN, 2023).

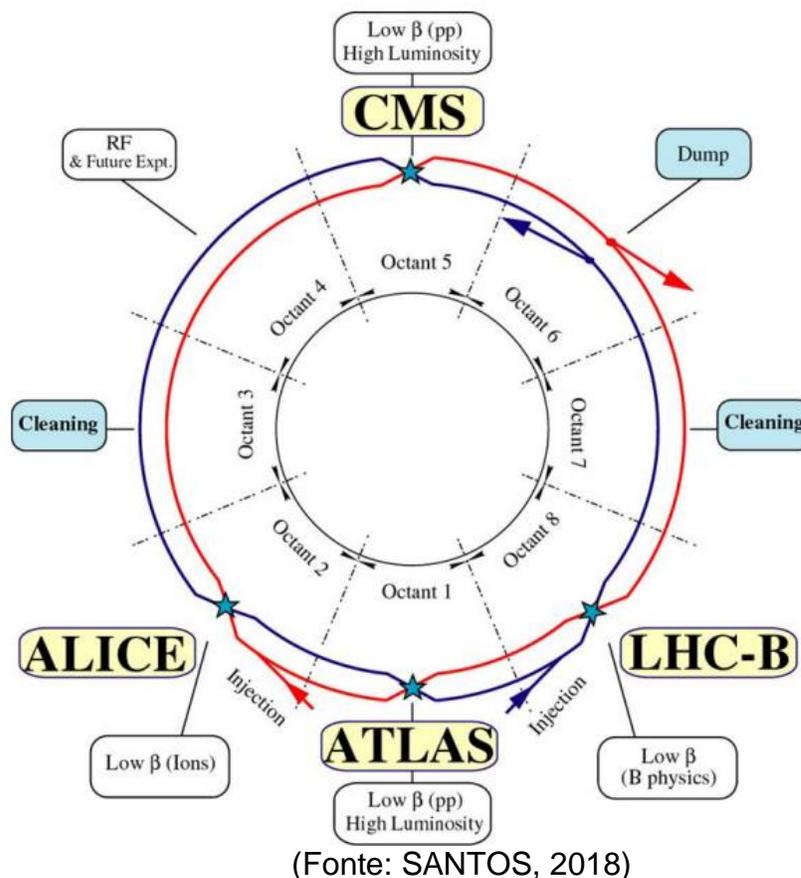
Figura 1.1 - Túnel do acelerador de prótons e antiprótons do CERN.



(Fonte: TIPLER; LIEWELLYN, 2017)

A Figura 1.2 descreve os sete experimentos instalados no LHC: A Large Ion Collider Experiment (ALICE), A Toroidal LHC ApparatuS (ATLAS), the Compact Muon Solenoid (CMS), the Large Hadron Collider beauty (LHCb), the Large Hadron Collider forward (LHCf), the TOTal Elastic and diffractive cross section Measurement (TOTEM) e o Monopole e Detector de Exóticos no LHC (MoEDAL). Dentre esses, o ALICE, o ATLAS, o CMS e LHCb são os maiores experimentos instalados ao redor dos quatro pontos de colisão LHC, sendo o ATLAS o de maior dimensão (CERN, 2023).

Figura 1.2 – Representação da disposição dos detectores de partículas ao longo de Lareg Hardron Collider.



O Experimento ATLAS é um detector de uso geral, que tem como uma das suas principais características seu enorme sistema de ímãs, constituído por oito bobinas magnéticas supercondutoras de 25 m de comprimento (CERN, 2017).

Ele é responsável por processar as informações provenientes das colisões próton-próton no LHC, por meio de um sistema de detecção baseado em dois níveis de filtragem online, de modo que, ao final do último nível, as informações sejam armazenadas definitivamente, possibilitando estudos posteriores (ATLAS Collaboration, 2008).

A detecção de partículas, no entanto, é um campo desafiador e complexo, e existem alguns fatores que podem dificultar a identificação e separação dos sinais das partículas de interesse. Dentre eles estão a presença de ruído de fundo na classificação dos eventos de interesse, que são sinais indesejados e interferências de outras fontes. Para tanto, é importante garantir a realização de uma medição precisa da energia das colisões, para uma melhor classificação das partículas resultantes. (ATLAS Collaboration, 2008).

Além disso, as altas taxas de eventos, demandam uma grande capacidade de processamento e aquisição de dados em tempo real. Diante disso, é necessário realizar a calibração de energia nos detectores para garantir a seleção mais precisa das partículas de interesse reduzindo assim a necessidade de armazenamento e processamento da informação (ATLAS Collaboration, 2008).

Uma das fontes de incerteza em sistemas de detecção está relacionada ao seu sistema de calibração. Isso porque, geralmente essa calibração envolve o ajuste e a correção dos parâmetros do detector, a fim de obter medidas precisas e consistentes das propriedades que se deseja estudar.

Diante disso a calibração de sistemas de detecção desempenha um papel fundamental na obtenção de resultados precisos e confiáveis em experimentos de física de altas energias. No Experimento ATLAS, o *High-Level Trigger* (HLT), é responsável pela seleção rápida e eficiente de eventos de interesse, exigindo um sistema de calibração robusto e eficaz para garantir a qualidade dos resultados obtidos.

Uma técnica de aprendizado de máquina que pode ser aplicada na calibração de sistemas de detecção são as árvores de decisão. A árvore de decisão é uma técnica de aprendizado de máquina amplamente utilizada na classificação e regressão de dados

(HASTIE et al., 2016). Sua aplicação na calibração de sistemas de detecção tem se mostrado promissora, permitindo a otimização dos parâmetros do detector e a correção de eventuais desvios sistemáticos (HASTIE et al., 2016).

Neste trabalho será analisado e discutido o impacto do uso de diferentes configurações de entrada para uma *Gradient Boosted Decision Tree* na etapa rápida do HLT, através da análise de dados simulados, afim de avaliar o desempenho e a eficácia dessa abordagem de calibração.

Espera-se que a implementação desse sistema de calibração contribua para melhorar a precisão na reconstrução de partículas e a identificação correta de eventos de interesse.

A física de partículas desempenha um papel fundamental no avanço do conhecimento científico e na compreensão fundamental da natureza. Ela tem uma grande importância não só no entendimento da estrutura do universo, mas também tem contribuído para avanços em áreas como a medicina. Atualmente, muitas das tecnologias empregadas no diagnóstico e tratamento de doenças, como a tomografia por emissão de pósitrons, responsável por identificar tumores e metástase em pacientes com câncer, surgiram a partir dos resultados de estudos e pesquisas em Física de Partículas (COSTA, 2018).

## **1.1. Motivação**

A física de partículas é uma área de pesquisa que busca compreender a estrutura fundamental da matéria e as interações entre as partículas elementares (PERATONI,2012). Nesse contexto, o Experimento ATLAS, desempenha um papel crucial na busca por novas partículas e fenômenos que possam responder a questões sobre a origem e a natureza do universo.

O experimento ATLAS gera uma grande quantidade de dados durante a colisão das partículas. Durante as colisões, grande parte da informação produzida já é conhecida (ruído de fundo) e, em muitos casos essas assinaturas de fundo tem um perfil

de deposição de energia similar ao evento de interesse (PERATONI, 2012). Diante disso, o *High-Level Trigger* (HLT), componente crucial do sistema de filtragem de dados do ATLAS, é importante para a realização eficiente da seleção dos eventos de interesse científico.

O HLT é composto por uma etapa rápida e uma etapa precisa. Devido à sua dependência das informações dos calorímetros, o sistema de trigger é suscetível a erros de medição de energia. Esses erros podem ocorrer durante o processamento online, uma vez que, nem toda a resolução do detector está disponível nesse estágio (SIMAS FILHO, et al., 2021). Por isso, o uso de técnicas de reconstrução de trajetórias de partículas, identificação de assinaturas características e tomada de decisão baseada em informações relevantes é essencial para obter medidas precisas das partículas produzidas nas colisões.

Atualmente, etapas de calibração, que melhoram a resolução das medições dos calorímetros, são realizadas apenas no processamento offline e na etapa precisa do HLT (SIMAS FILHO, et al., 2021). Dessa forma, com o objetivo de reduzir os requisitos computacionais e melhorar a seleção de partículas eletromagnéticas foi realizado um estudo da implementação de sistema de calibração na etapa rápida do HLT, foi realizado um estudo da implementação de um sistema de calibração utilizando Árvore de Decisão com Reforço por Gradiente (*Gradient Boosted Decision Tree*).

Para realizar esse estudo foi utilizado um regressor baseado em *Gradient Boosted Decision Trees*, treinado com dados simulados, e dois cenários de calibração foram investigados, com o objetivo de analisar os efeitos de cada cenário no desempenho do sistema.

Um sistema de filtragem eficiente é fundamental para extrair os eventos mais relevantes da grande quantidade de dados gerados pelo Experimento ATLAS. Uma etapa rápida do HLT otimizada contribuirá para uma seleção mais precisa, diminuindo a taxa de eventos a serem processados na etapa posterior e permitindo uma análise mais eficiente dos fenômenos físicos de interesse.

Compreender e melhorar a fase rápida do HLT é de extrema importância para avançar na área da física de partículas e possibilitar a descoberta de novas partículas, fenômenos e processos fundamentais que são essenciais para a compreensão do Universo. Desse modo, este trabalho tem como objetivo fornecer novas perspectivas e soluções para aprimorar a seleção e a eficiência da etapa rápida do HLT no contexto do Experimento ATLAS.

## **1.2. Objetivos**

### **1.2.1. Objetivo Geral**

O objetivo geral deste trabalho é investigar se, utilizando um número menor de entradas para a *Gradient Boosted Decision Tree* (GBDT) na calibração da etapa rápida do HLT, é possível melhorar a estimação da energia das partículas e reduzir a demanda de processamento computacional, de modo adequado.

### **1.2.2. Objetivos Específicos:**

- Realizar uma revisão bibliográfica sobre o funcionamento do HLT, e algumas das técnicas de aprendizado de máquina utilizadas na seleção de partículas;
- Analisar os dados simulados do Experimento ATLAS, avaliando a performance da GBTD para os cenários de calibração propostos;
- Comparar os resultados obtidos na estimação de energia com e sem a implementação da GBTD em cenários com entradas variadas;
- Avaliar o erro relativo da estimação da energia após a aplicação da calibração utilizando a GBTD etapa rápida do HLT.

### 1.3. Organização do Trabalho

No Capítulo 1 é apresentada uma visão geral do LHC, destacando importância do Experimento ATLAS, as dificuldades relacionadas a detecção de partículas e a importância da etapa rápida do HLT nesse contexto. Também são apresentados os objetivos e a relevância do estudo realizado, além de uma breve descrição da estrutura geral do trabalho.

No Capítulo 2 é feita uma revisão da literatura relacionada à etapa rápida do HLT, abordando conceitos fundamentais, técnicas e abordagens utilizadas nessa etapa. É explorada a teoria relacionada as árvores de decisão, discutindo seus princípios e alguns tipos de implementação.

No capítulo 3 a metodologia adotada para realizar a análise da calibração proposta é realizada. Com isso é detalhado o processo de implementação da GBTD e o formato dos dados utilizados. Além disso, também é discutido o ajuste dos parâmetros das árvores de decisão.

No capítulo 4 é apresentado os resultados obtidos por meio da aplicação do sistema de calibração proposto, através de gráficos que representam dispersão da medição e do erro associado as medições. Além disso, também são realizadas análises comparativas dos resultados do sistema proposto com a abordagem de calibração já existente.

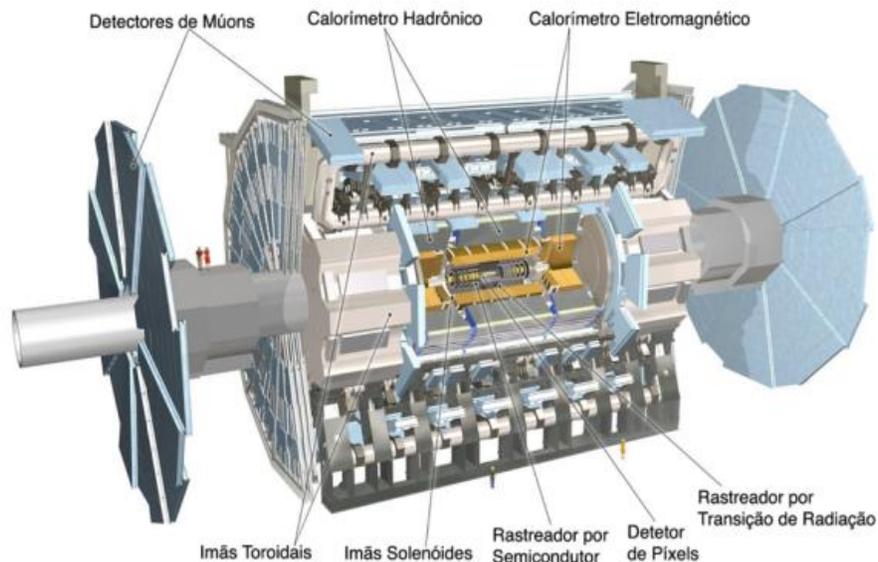
No Capítulo 5 os principais pontos abordados ao longo do trabalho são pontuados reiterando a importância da etapa rápida do HLT e a contribuição do sistema de calibração proposto. Além disso, as conclusões obtidas em relação aos objetivos definidos são apresentadas, e sugeridas possíveis direções para pesquisas futuras nessa área de estudo.

## 2. FUNDAMENTAÇÃO TEÓRICA

### 2.1. Experimento ATLAS

Construído em formato cilíndrico, com raio de 12,5 m, comprimento de 44 m e com massa de aproximadamente 7.000 toneladas, o ATLAS é o maior detector de propósito geral do LHC. Ele processa as informações oriundas das colisões próton-próton e sua estrutura, é dividida em subdetectores que possuem funções específicas, sendo eles o Detector Interno (ID), o Calorímetro Eletromagnético (EMEC), o Calorímetro Hadrônico (HEC), o Calorímetro *Forward* (FCal) e o Espectrômetro de Múons (ATLAS Collaboration, 2008).

Figura 2.1 – Vista em corte do Experimento ATLAS e seus respectivos subdetectores.



(Fonte: TORRES, 2010)

Como mostra a Figura 2.1, o ID está localizado na parte mais próxima do ponto de colisão e é responsável por identificar a trajetória das partículas carregadas eletricamente. Ele é composto pelo Detetor de Pixels, Rastreador por Semicondutor e Rastreador por Transição de Radiação. O EMEC, por sua vez, detecta e absorve elétrons,

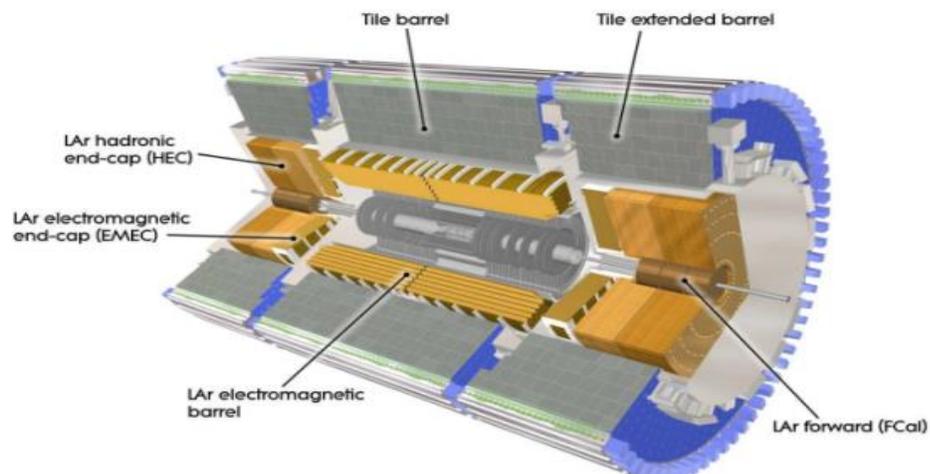
fótons e pósitrons, e o HEC possibilita a análise de deposição energética de partículas hadrônicas (ATLAS Collaboration, 2008).

Localizado no ponto mais afastado do centro das colisões, o FCal suporta altas taxas de radiação e fluxo de partículas, e é usado para estudo de processos que possuem energia transversa faltante. Como os Múons conseguem se propagar para além dos calorímetros, a sua detecção é feita pelo Espectrômetro de Múons (ATLAS Collaboration, 2008).

O sistema de calorimetria do ATLAS é fundamental na medição da energia das partículas. Na física de partículas, calorimetria é o estudo que permite a detecção e a medição de propriedades das partículas a partir da absorção das mesmas por um bloco de matéria (WIGMANS, 2000).

Embora existam vários tipos de calorímetros, todos são baseados em um processo de detecção destrutivo, sendo assim, após a interação com as partículas, não é possível realizar nenhuma outra medição sobre as mesmas. Além disso, dependendo da sua composição, os calorímetros podem ser homogêneos ou de amostragem. No primeiro apenas um tipo de material sensível a partícula incidente é utilizado, enquanto que no segundo são dois, um material passivo absorve a energia da partícula, e um ativo, amostrador gera um pulso de luz para a sua detecção (WIGMANS, 2000).

Figura 2.2 – Diagrama ilustrativo do sistema de calorimetria do Experimento ATLAS.



(Fonte: ATLAS Collaboration, 2008)

A Figura 2.2 mostra a estrutura do sistema de calorimetria do ATLAS que se subdivide em eletromagnético e hadrônico. O calorímetro eletromagnético está localizado na parte mais interna, e é formado por uma estrutura sequencial de argônio líquido, e placas de chumbo no padrão de acordeons. O hadrônico, por sua vez, identifica partículas como os nêutrons, píons, e prótons, está posicionado na parte mais externa, sendo composto por telhas de material cintilante (ATLAS Collaboration, 2008).

As informações oriundas das colisões detectadas pelo Experimento ATLAS, são processadas em um sistema (*trigger*) que possui dois níveis de filtragem dos eventos de interesse. O primeiro nível L1, é implementado em hardware dedicado e usa um subconjunto de informações do detector para reduzir a taxa de eventos a um valor de projeto de no máximo 100 kHz (ATLAS Collaboration, 2008).

O segundo, High-Level Trigger (HLT), é baseado em software e se subdivide em uma etapa rápida, e uma precisa. Cada nível do sistema de *trigger* refina as decisões tomadas no nível anterior e, quando necessário, aplica critérios de seleção adicionais. O L1 utiliza uma quantidade limitada da informação total do detector para tomar uma decisão em menos de 2,5  $\mu$ s, e o HLT, acessa mais informações do detector para uma taxa final de até 200 Hz (ATLAS Collaboration, 2008).

## 2.2. Sistema de Seleção Online

Com o objetivo de realizar novas descobertas, o LHC opera há uma alta taxa de colisão, e tem como um dos seus objetivos a identificação do bóson de Higgs. A identificação do bosón de Higgs, não pode ser feita diretamente, uma vez que este decai para partículas mais estáveis antes que o sistema de detecção seja capaz de identificá-lo. Diante disso, para caracteriza-lo é necessário detectar as partículas resultantes do decaimento, teoricamente presumida. Dessa forma, a seleção dos dados tem que garantir que estes eventos de interesse não sejam perdidos e que a taxa de saída seja reduzida para um valor que seja viável o armazenamento para estudos posteriores (NESSI, Marzio et al, 2003).

O Experimento ATLAS foi projetado para observar até 1,7 bilhão de colisões próton-próton por segundo, com um volume de dados combinado de mais de 60 milhões de megabytes por segundo. Entretanto, apenas alguns desses eventos conterão características interessantes que podem levar a novas descobertas. Sendo assim, para reduzir o fluxo de dados a níveis gerenciáveis, o ATLAS utiliza um sistema especial de seleção de eventos - o *trigger* - que seleciona os eventos com características distintas para análises físicas (CERN, 2023).

Esse sistema de acionamento e aquisição de dados garante condições ideais de coleta de dados e seleciona os eventos de colisão mais interessantes para serem explorados. Ele é composto por uma camada desenvolvida em hardware dedicado (L1) que realiza a primeira etapa de seleção dos eventos, seguido por uma segunda camada de software (*High-Level-Trigger*), em que operam os discriminadores, responsáveis por realizar a classificação dos eventos registrados pelo detector (CERN, 2023).

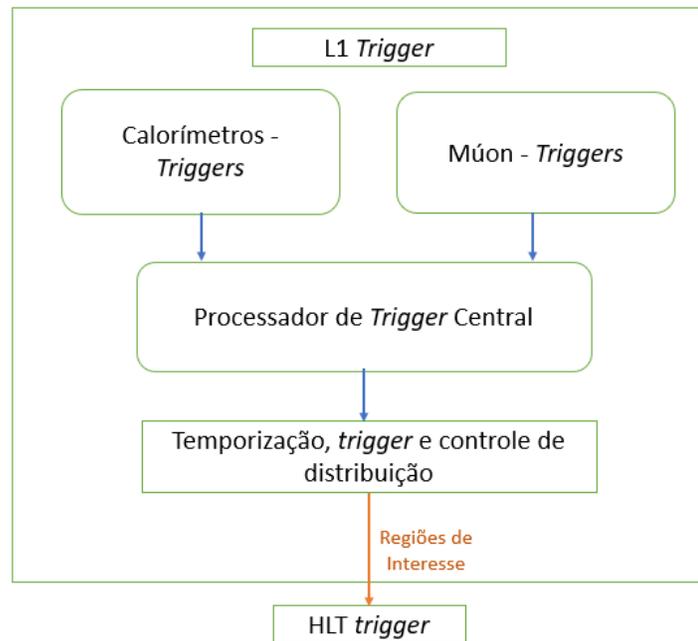
### 2.3. Primeiro Nível de Filtragem

O primeiro nível de filtragem, L1, busca por múons, elétrons, fótons, jatos e  $\tau$ -leptons de alto momento transverso que decaem em hádrons, bem como energia transversal total ( $E_t$ ) e energia perdida (*missing*  $E_t$ ). Ele é responsável pela primeira avaliação e seleção preliminar dos eventos, além de identificar das regiões de interesse (RoI), baseadas nas informações dos calorímetros e detectores de múons (ATLAS Collaboration, 2008).

As Rols indicam as posições do detector em que seu processo de seleção identificou eventos potencialmente interessantes. Os resultados obtidos são processados pelo processador de *trigger* central, que implementa um "menu" de acionamento composto de combinações de seleções de acionamento. Após passarem pela camada L1, os eventos são transferidos para os próximos estágios de filtragem do detector e, posteriormente, para a aquisição de dados (ACHENBACH, 2008).

Figura 2.3 Diagrama de blocos que representa as etapas que ocorrem no processo de filtragem do sistema de trigger L1 do detector ATLAS.

(Fonte: Adaptado ATLAS Collaboration,2008)



Conforme a Figura 2.3, nota-se que a decisão geral de aceitação de L1 é tomada pelo Central Trigger Processor (CTP), que é responsável por tratar as informações oriundas dos triggers dos sistemas de calorimetria e de múons para produzir o *trigger* de decisão do nível L1.

Os critérios de seleção são previamente definidos através de um menu de *trigger* podem ser programados com até 256 itens distintos, sendo cada item uma combinação de requisitos dos dados de entrada. Além disso, a decisão de disparo, juntamente com o *clock* de 40,08 MHz e outros sinais, é distribuída para os sistemas de leitura do detector por meio do sistema de temporização, disparo e controle (TTC). Após serem aceitas pelo L1 essas informações são enviadas como Rols, para a camada seguinte HLT (ATLAS Collaboration, 2008).

Um das principais funções do *trigger* L1 é a identificação não ambígua do cruzamento de feixe de interesse. Para que isso aconteça é importante que a latência do sistema seja pequena, ou seja, o tempo gasto para gerar e distribuir o resultado da

decisão precisa ser o menor possível, uma vez que o intervalo entre as colisões é muito curto (ACHENBACH, 2008).

Entre o primeiro e o segundo nível de filtragem online do ATLAS existem alguns módulos de armazenamento. Dentre eles estão dois sistemas de memórias, um que guarda as informações oriundas das colisões, e outro que armazena somente as informações de interesse. Os *Readout Drivers* (ROD) são os responsáveis por coletar as informações de saída do L1 para os *ReadOut Buffers* (ROB), que por sua vez, armazenam os dados até que o evento seja validado ou descartado pelo próximo nível (ATLAS Collaboration, 2008).

O L1 *trigger*, portanto, recebe as informações dos calorímetros e dos detectores de múons e, existindo a ocorrência de algum evento de interesse da pesquisa do ATLAS, é gerado um trigger de decisão, que é direcionado ao HLT como regiões de interesse (ATLAS Collaboration, 2008).

## **2.4. Segundo Nível de Filtragem**

Após o primeiro nível de filtragem L1 encontra-se o nível mais elevado do sistema de trigger o HLT (*High-Level Trigger*). Ele tem como ponto de partida os resultados provenientes da camada L1, que são as informações sobre o tipo de trigger e regiões de interesse (Rols). Diante disso, o HLT, refina a seleção de eventos realizada na etapa anterior (ATLAS Collaboration, 2008).

A finalidade do *High-Level-Trigger* do ATLAS é validar a decisão do nível anterior, reduzindo a taxa de eventos de 75 kHz para aproximadamente 3 kHz, num tempo de processamento de no máximo 40 ms. Pra isso, ele opera em duas etapas a rápida e a precisa. Na primeira, ele manipula as informações das regiões de interesse selecionadas pelo L1, e em seguida os eventos de interesse são utilizados na etapa precisa, para refinar a decisão final do HLT (SIMAS FILHO, et al., 2021).

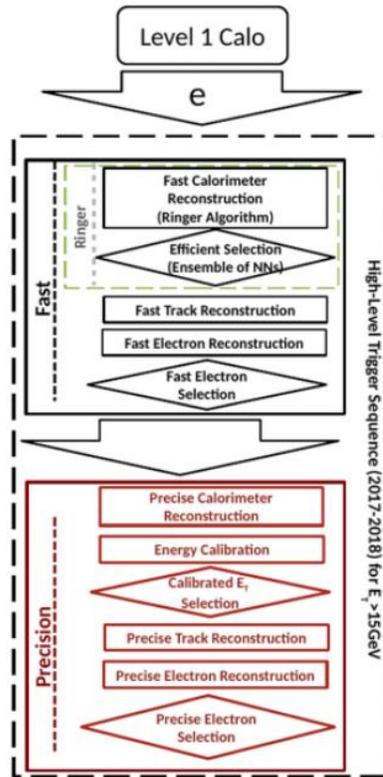
Quando um evento é aceito pelo L1, as suas informações ficam armazenadas nos *ReadOut Buffers* (ROB), pertencentes ao sistema de leitura do detetor, o *ReadOut*

*System* (RO). O sistema de leitura, recebe e armazena temporariamente os dados em *buffers* locais, e posteriormente, o HLT solicita os dados do evento associados às Rols. Os eventos selecionados pelo acionador L2 são então transferidos para o sistema de criação de eventos e, subseqüentemente, para o filtro de eventos para seleção final. Os eventos selecionados pelo filtro de eventos são transferidos para o armazenamento permanente (ATLAS Collaboration, 2008).

Os principais elementos, do segundo nível de filtragem, são os construtores das regiões de interesse – *RoI Builder* (RoIB), os supervisores – L2SV, os processadores – L2P, e os sub-sistemas de leitura – ROS (Read-Out Sub-system). Os construtores das regiões de interesse (RoIB) possuem a função de agrupar as informações, provenientes L1 trigger e, transmiti-las para os supervisores do nível 2 (L2SV). Os supervisores, por sua vez, selecionam os processadores (L2P) e enviam as informações provenientes do L1 para serem processadas. Por meio da execução de softwares de seleção de eventos, e das informações dos subsistemas de leitura, os processadores obtém os resultados, que irão formar o *trigger* de decisão do HLT (NESSI, Marzio et, 2003).

Primeiramente o HLT começa o seu processo de filtragem com as Rols fornecidas pelo nível L1. Ele aplica decisões de acionamento em uma série de etapas, como mostra a Figura 2.4, cada uma refinando as informações existentes por meio da aquisição de dados adicionais de um número cada vez maior de detectores. Uma lista de assinaturas físicas (cadeias de acionamento), reconstrução de eventos implementada (extração de características) e algoritmos de seleção são usados para criar tabelas de assinaturas e sequências para todas as etapas do HLT (ATLAS Collaboration, 2008).

Figura 2.4 – Diagrama que representa o fluxo de dados no segundo nível de filtragem online do Experimento ATLAS.



(Fonte: SIMAS FILHO, et al., 2021)

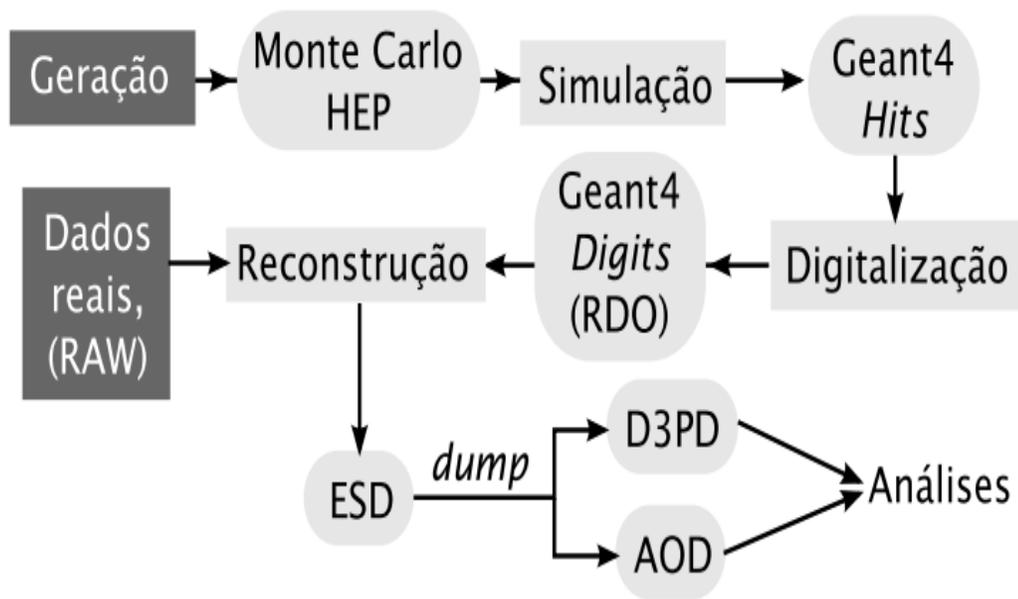
Os algoritmos de extração de características normalmente solicitam dados do detector de dentro da RoI e tentam identificar características, por exemplo, um rastro ou um conjunto de calorímetros. Em seguida, um algoritmo de hipótese determina se o recurso identificado atende aos critérios necessários para continuar (ATLAS Collaboration, 2008).

A decisão de rejeitar o evento ou continuar baseia-se na validade das assinaturas. Desse modo, os eventos podem ser rejeitados logo após uma etapa intermediária se nenhuma assinatura permanecer viável. Diante disso, o conjunto completo de dados associado à RoI é transferido somente para os eventos que atendem aos critérios completos de seleção do HLT (ATLAS Collaboration, 2008).

Afim de garantir o bom funcionamento do detector e diminuir o tempo de desenvolvimento, é necessário o uso de ferramentas que possam confirmar e testar todas as funcionalidades dos seus diversos constituintes. Diante disso, o ATLAS possui um framework, o ATHENA, em que é possível testar os algoritmos em um ambiente que emula o funcionamento do sistema de filtragem do detector (PERANTONI,2012).

Esse framework possui uma estrutura de funcionamento que possibilita que os algoritmos sejam testados de modo online ou offline. O modo online é o que apresenta características mais parecidas com o real funcionamento do sistema de trigger do HLT, uma vez que os seus componentes estão em plena operação. Por outro lado, no modo offline é possível usar somente as aplicações específicas de uma parte do sistema de filtragem. Além disso, o ATHENA possibilita ainda, a realização de simulações de Monte Carlo das partículas produzidas pelos subdetectores do ATLAS depois das colisões no LHC, como os eventos de elétrons e jatos, no sistema de calorimetria (PERANTONI,2012).

Figura 2.5 – Diagrama de blocos que representa a geração de dados do ATLAS para análise via ATHENA



(Fonte: CIODARO, 2012)

A Figura 2.5 mostra as etapas de transformação dos dados no ATLAS. Todos esses processos podem ser realizados utilizando o Athena, sendo que os dados finais podem ser obtidos a partir de simulações de Monte de Carlo, ou pelos detectores de colisões.

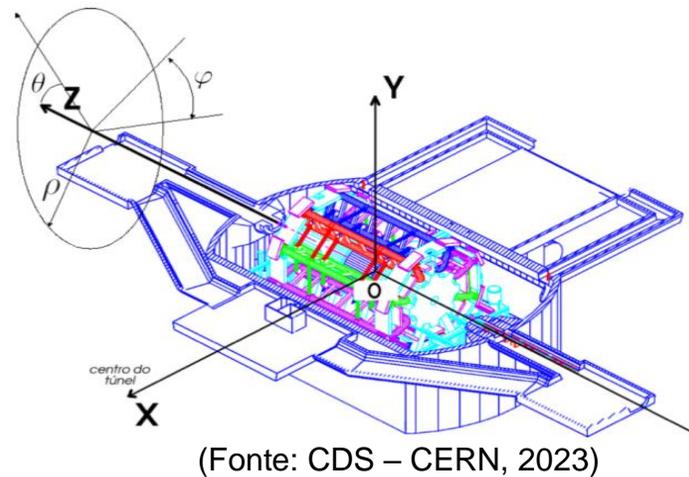
## **2.5. Identificação dos Elétrons**

O sistema de calorímetros do ATLAS é composto por 187.652 sensores e tem como objetivo medir a energia depositada pelas geradas nas colisões do LHC (ATLAS Collaboration, 2008). Quando um evento candidato a elétron é identificado pelo nível L1, o HLT usa a granularidade do detector com a finalidade de validar se o evento é um elétron ou não (NESSI, Marzio et al, 2003).

Essa classificação é muito importante, uma vez que os elétrons de momento elevado representam uma das principais características do Bosón de Higgs, que é uma partícula fundamental prevista pelo Modelo Padrão da física de partículas. (ANJOS, 2006). Além disso, visto que jatos de outras partículas também excitam o calorímetro eletromagnético, a energia armazenada pelo calorímetro hadrônico também é mensurada, e utilizada na discriminação das partículas de elétrons (ATLAS Collaboration, 2008).

A identificação de elétrons no ATLAS é baseada em informações provenientes de diferentes partes do detector, e por isso para reconstruir a posição e energia das partículas é utilizado o sistema de coordenadas cilíndricas. Desse modo é possível descrever a posição e direção das partículas produzidas no LHC que são informações importantes para a identificação das partículas geradas nas colisões nele acontecem (ANJOS,2006).

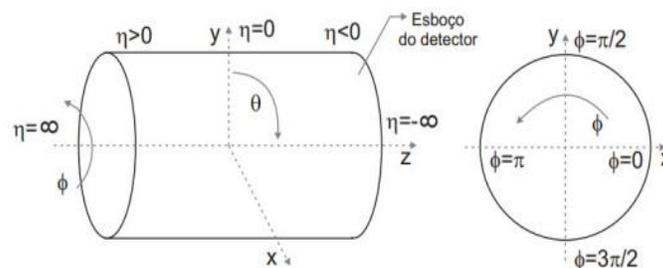
Figura 2.6 – Representação do sistema de coordenadas do Experimento ATLAS.



Na Figura 2.6 o eixo z aponta na direção do feixe de partículas e o eixo x aponta horizontalmente para fora do ponto de origem, com valores positivos apontando para o centro do anel do LHC. O eixo y, por sua vez, aponta verticalmente para cima, com valores positivos apontando para o topo do detector.

No contexto do ATLAS, a "pseudo-rapidez" ( $\eta$ ) é uma medida importante para descrever a direção de movimento das partículas produzidas nas colisões de alta energia. Ela é definida como uma função do ângulo polar da partícula em relação à direção do feixe de partículas (ATLAS Collaboration, 2008). A Figura 2.7 descreve  $\eta$  no sistema de coordenadas cilíndricas.

Figura 2.57 – Diagrama do sistema de coordenadas cilíndricas do ATLAS.



As coordenadas que indicam a posição onde ocorrem as colisões podem ser definidas a partir das seguintes equações:

$$\phi = \arctan\left(\frac{y}{x}\right) \quad (2.1)$$

$$\theta = \arccos\left(\frac{z}{R}\right) \quad (2.2)$$

$$\eta = -\ln\left(\tan\frac{\theta}{2}\right) \quad (2.3)$$

$$E_T = E_{\sin}(\theta) \quad (2.4)$$

A Equação 2.1 indica a rotação  $\phi$  em torno do eixo de colisão (z); o ângulo polar  $\theta$  definido pela Equação 2.2, em que R é a distância das partículas do ponto de colisão até os sensores). A pseudo-rapidez  $\eta$ , que corresponde direção de propagação das partículas após a colisão, é descrita na Equação 2.3 em função de  $\theta$ , e a energia transversa  $E_T$ , definida pela Equação 2.4 (ATLAS Collaboration, 2017).

A energia transversa, é uma quantidade importante medida no Experimento ATLAS. Ela é a componente da energia de uma partícula que está perpendicular à direção do feixe de partículas, que nesse caso, é ao longo do feixe circular do LHC, e é um indicador da energia total de uma partícula produzida nas colisões. Além disso, a energia transversa é usada para definir a energia transversa ausente, que é uma medida da energia não detectada no detector, que auxilia na identificação de partículas neutras como neutrinos (ATLAS Collaboration, 2008).

Atualmente para a identificação online dos elétrons, o experimento ATLAS utiliza como um dos seus discriminadores o Neural Ringer, em que o perfil de deposição de energia é utilizado como entrada para uma rede neural de múltiplas camadas, que atua como um classificador (SIMAS FILHO, et al., 2021).

Nesse modelo, os dados obtidos pelos sensores dos calorímetros são dispostos em um vetor de 100 posições, que correspondem a intensidade da energia depositada nas camadas do detector ao longo do caminho descrito pela partícula. Essas informações são combinadas por meio de algoritmos complexos, e utilizadas como entrada do sistema de classificação baseado em rede neural artificial, a fim de identificar

com precisão os elétrons produzidos nas colisões de partículas (SIMAS FILHO, et al., 2021).

## **2.6. Neural Ringer**

Após serem classificadas no nível de filtragem L1, as informações de calorimetria das Rols são compactas no formato de anéis concêntricos de deposição de energia, para serem posteriormente processadas por uma rede neural. Essa rede neural é então treinada usando um grande conjunto de dados de eventos simulados, e com isso aprende a reconhecer os padrões associados as assinaturas de elétrons. Por meio desse processo, a rede neural torna-se capaz de distinguir os elétrons do ruído de fundo e de outras partículas (WIGMANS, 2000).

No ATLAS o algoritmo com base em rede neural projetado para melhorar a identificação dos elétrons em eventos de colisão de alta energia é o Neural Ringer. Quando os elétrons depositam energia nos calorímetros, geralmente anéis de energia são formados. Diante disso, o Neural Ringer utiliza esses clusters de energia como entrada e emprega uma rede neural para classificar os elétrons (ANJOS, 2018).

O Neural Ringer realiza dois procedimentos a montagem dos anéis e a classificação do evento. Nessa etapa, primeiramente o algoritmo detecta a célula de maior deposição de energia, que formará o primeiro anel. A energia das células adjacentes ao primeiro anel forma o segundo anel (CIODARO, 2012).

Esse processo é sucessivamente repetido até que os cem anéis (ou um número máximo) da região de interesse sejam utilizados, ao longo das camadas do calorímetro. Os dados energéticos gerados em cada anel são concatenados e adicionados em um vetor de cem posições, que posteriormente será a entrada do classificador neural (CIODARO, 2012).

Tabela 2.61 – Número de anéis por camada do sistema de calorimetria do ATLAS.

Camada	PS	EM1	EM2	EM3	HAD0	HAD1	HAD2
Anéis	08	64	08	08	04	04	04

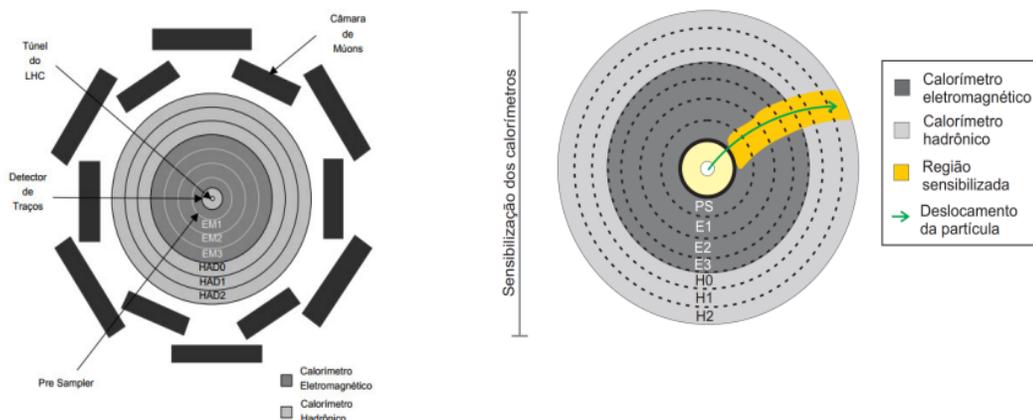
(Fonte: SIMAS FILHO, et al., 2021).

A Tabela 2.1 apresenta a quantidade de anéis por camada, de acordo com a sua granularidade. As camadas do calorímetro eletromagnético são descritas por EM1, EM2 e EM3. A do calorímetro hadrônico são a HAD0, HAD1 E HAD2, e a do pré-amostrador, por sua vez é definida por PS.

O pré-amostrador é uma fina camada separada de argônio líquido, localizado mais próximo da região de interação dos feixes de partículas. Ele é composto por 64 setores azimutais idênticos e sua principal função é medir a energia depositada por partículas carregadas antes que elas se espalhem ou interajam com os materiais do detector (ATLAS Collaboration, 2008).

A Figura 2.8 representa o arranjo de anéis para as camadas dos calorímetros. Essa distribuição em anéis concêntricos conserva as características de deposição de energia das partículas, possibilitando assim a sua classificação (WINGMANS, 2000).

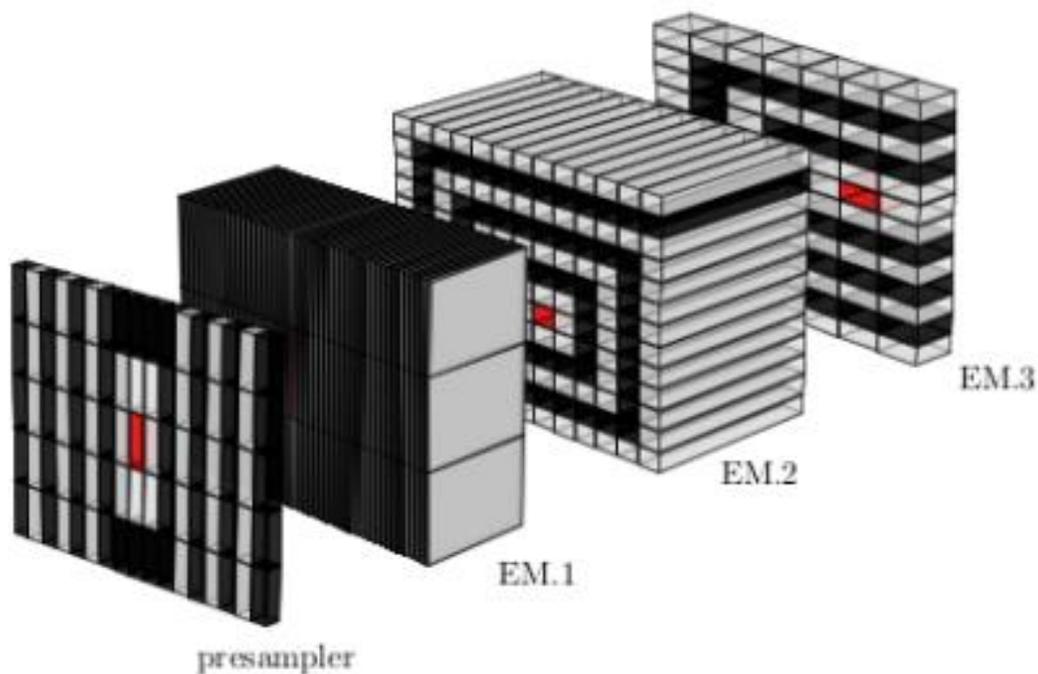
Figura 2.68 – Diagrama de um corte transversal do detector que representa as suas camadas e as possíveis trajetórias de um evento.



(Fonte: SIMAS FILHO, 2010)

O Neural Ringer atua na etapa rápida e é um algoritmo que combina extração de características e teste de hipótese para a discriminação de elétrons. A informação de calorimetria das regiões de interesse, selecionadas pelo L1, é compactada na forma de anéis concêntricos de deposição de energia como mostra a Figura 2.9 (SIMAS FILHO, et al., 2021).

Figura 2.69 – Representação dos anéis e suas diferentes granularidades de quatro camadas do calorímetro. Os pontos em vermelho, são as células mais energéticas.



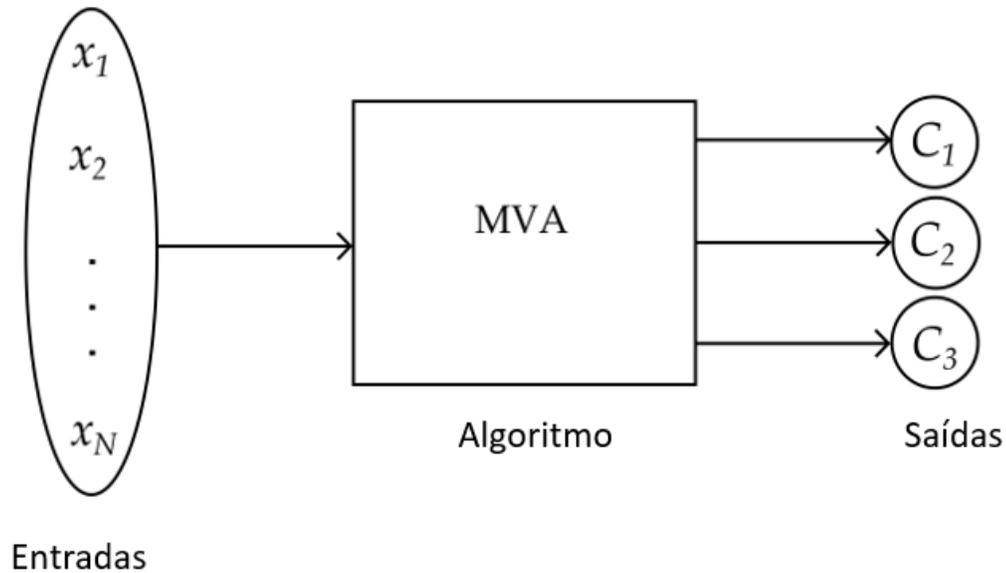
(Fonte: SIMAS FILHO, et al., 2021)

## 2.7. Algoritmos de Aprendizado de Máquina

O aprendizado de máquina (do inglês: *Machine Learning*) utiliza várias técnicas com o objetivo de extrair informações significativas de um conjunto de dados. Um algoritmo de aprendizado de máquina é um modelo matemático que contém parâmetros que são otimizados para que o modelo possa descrever os dados de forma eficiente. A fase em que são encontrados os melhores valores para os parâmetros é chamada de treinamento, e é nessa etapa que ocorre o aprendizado do algoritmo a

partir dos dados fornecidos. Após essa fase de aprendizado, o algoritmo é capaz de realizar previsões sobre dados desconhecidos (ALPAYDIN, 2014).

Figura 2.10 – Diagrama de um algoritmo de aprendizado de máquina. Nesse caso, são fornecidos N recursos de entrada, e o método multivariado (MVA) atribui cada evento da amostra a uma das três classes.



(Fonte: Adaptado ALPAYDIN, 2014)

Os algoritmos recebem como entrada um conjunto de variáveis para cada ponto de dados, como mostra a Figura 2.10. Essas variáveis são chamadas de recursos e contêm informações sobre as propriedades de cada conjunto de dados. Esses recursos podem ser, por exemplo, a energia transversa e a pseudo-rapidez das partículas, em uma colisão, em que cada grupo de dados representa um único evento detectado no experimento.

Os algoritmos de *machine learning* podem implementar dois tipos de aprendizado: supervisionado ou não supervisionado. O primeiro requer que o conjunto de dados seja rotulado com o resultado desejado que o algoritmo deve produzir para os valores de entrada desse evento específico (MURPHY, 2012).

Desse modo, o objetivo do aprendizado supervisionado é aprender a relação entre a entrada e saída. Ele aprende ajustando seus parâmetros a fim de minimizar o

erro entre sua saída e o rótulo. Uma vez que o modelo tenha sido treinado, ele pode ser usado para fazer previsões ou classificações em novos exemplos em que o rótulo real não é conhecido (MURPHY, 2012).

O aprendizado não supervisionado, por outro lado, é usado quando não há rotulagem disponível para os dados. Os eventos são agrupados com base nas semelhanças de seus recursos, e são frequentemente usados para pré-processar os dados antes de usar algoritmos supervisionados. Dessa forma, o aprendizado não supervisionado busca identificar similaridades, diferenças e estruturas subjacentes nos dados sem ter conhecimento prévio dos grupos ou classes (KUHN; JOHNSONL, 2013).

Dois aspectos importantes a serem considerados na criação de um algoritmo de aprendizado de máquina supervisionado são o viés (bias), ou erro de predição, e a variância. O primeiro é a diferença entre a saída do método MVA para um determinado conjunto de dados e o rótulo do evento. Ele pode indicar o quão bem o algoritmo compreendeu as conexões e correlações subjacentes nos dados (ALPAYDIN, 2014).

A variância é um conceito estatístico utilizada para representar a sensibilidade do algoritmo ao ruído aleatório na amostra de dados. Um classificador que tenha uma variância muito alta se ajustará perfeitamente aos dados de treinamento, não conseguindo generalizar quando receber um novo conjunto de dados (ALPAYDIN, 2014).

O erro de classificação de um algoritmo pode ser descrito pela Equação 2.5:

$$E = b^2 + V + \sigma_i \quad (2.5)$$

em que  $b$  é o erro de viés,  $V$  é a variância e  $\sigma_i$  é o chamado erro irreduzível, que se origina de propriedades intrínsecas dos dados e não pode ser melhorado (ALPAYDIN, 2014).

## 2.8. Árvores de Decisão

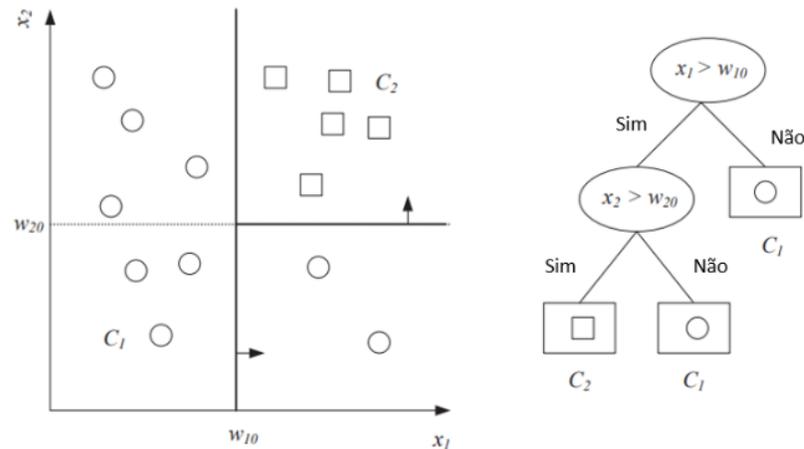
A indução de árvores de decisão é uma das formas mais simples e bem-sucedidas de aprendizado de máquina. Uma árvore de decisão representa uma função que recebe como entrada um vetor de valores de recursos e retorna uma "decisão" - um único valor de saída. Os valores de entrada e saída podem ser discretos ou contínuos. Ela chega à sua decisão executando uma sequência de testes (RUSSEL; NORVIG, 2020).

Para cada tarefa de classificação ou regressão, é possível escolher um algoritmo aprendizado de máquina diferente, e com o aumento da complexidade dos dados, os modelos mais complexos tendem a superar os mais simples, porque possuem uma maior capacidade de adaptação. As árvores de decisão, no entanto, iteram regras simples para criar limites de decisão, e ainda apresentar um bom desempenho em comparação com, por exemplo, perceptrons de múltiplas camadas testes (RUSSEL; NORVIG, 2020).

Uma árvore de decisão é um algoritmo de aprendizado supervisionado, que funciona operando repetidamente cortes no conjunto de dados, separando-o em subconjuntos menores que exibem uma melhor separação de classes do que o conjunto de dados original. Ao final do processo de treinamento, o espaço de recursos é dividido em regiões menores, e os dados não vistos anteriormente são classificados de acordo com a região em que se enquadram (ALPAYDIN, 2014).

As árvores de decisão também podem ser representadas esquematicamente através de um diagrama, com bolhas conectadas por linhas. As bolhas são chamadas de nós e representam subconjuntos da amostra, enquanto as linhas que conectam as bolhas mostram o resultado das diferentes decisões (ALPAYDIN, 2014).

Figura 2.11 – Representação de um conjunto de dados e a sua árvore de decisão correspondente. As elipses ou bolhas são os nós de decisão e os retângulos são os nós de folha. O nó de decisão se subdivide até que as classes estejam separadas.



(Fonte: Adaptado ALPAYDIN, 2014)

Na árvore de decisão representada na Figura 2.11 é possível perceber que cada nó contém um determinado subconjunto e o critério de corte que foi escolhido para esse subconjunto. Após cada corte, a subamostra é dividida em dois conjuntos menores, conectados ao original por linhas. O primeiro nó contém o conjunto completo de dados, e os nós finais (chamados de nós folha) correspondem às regiões em que o espaço de atributos está dividido.

Vários classificadores, chamados de aprendizes básicos, são treinados em diferentes subconjuntos de toda a amostra de treinamento. Na fase de aplicação, os eventos são classificados de acordo com a média ponderada das respostas. Um dos métodos em que vários classificadores são combinados é o *boosting*, que quando aplicado pode reduzir o erro de treinamento e validação, tornando os algoritmos mais robustos (KUHN; JOHNSONL, 2013).

Árvore *Boosting* é a combinação de modelos simples, para criar um modelo preditivo mais robusto e mais preciso. Ao invés de criar um novo modelo do zero, a técnica do *boosting* constrói iterativamente novos modelos para corrigir os erros cometidos pelos modelos anteriores. Dessa forma, cada novo modelo é treinado com os

resíduos dos modelos anteriores, com o objetivo de minimizar o erro (HASTIE et al., 2016).

Um dos principais algoritmos *boosting* é o *Gradient Boosting*, também conhecido como *Gradient Boosting Machine* (GBM) que é um algoritmo que faz a classificação e regressão através do modelo aditivo (HASTIE et al., 2016). Assim, esse sistema faz uso de modelos aditivos para gradativamente encontrar o melhor modelo aproximado.

Dado um conjunto de dados  $(x, y)_{i=1}^N$ , em que  $x = x_1, \dots, x_d$  refere-se às variáveis de entrada explicativas e  $y$  aos rótulos correspondentes da variável de resposta, o objetivo do GBM é reconstruir a dependência funcional desconhecida  $x \xrightarrow{f} y$  com a estimativa  $F(x)$  de modo que a função de perda especificada  $L(y_i, F(x))$  seja minimizado (HASTIE et al., 2016).

Dessa forma tem-se:

$$\hat{f}(x) = y \quad (2.6)$$

$$F_0(x) = \arg \min \sum_{i=1}^n L(y_i, \gamma) \quad (2.7)$$

Em que  $y_i$  é o valor observado e  $\gamma$  são os valores preditos.

Ao aplicar a Equação 2.7 tem-se o objetivo de encontrar valor predito que minimize esta soma. O passo seguinte é o processo de iteração de todas as árvores, iniciando com a primeira árvore  $M = 1$ .

Com  $M$  etapas de iteração, as estimativas de parâmetros podem ser escritas na forma incremental conforme a Equação 2.8 (HASTIE et al., 2016):

$$F(x) = \sum_{i=1}^M F_i(x) \quad (2.8)$$

Dentro desta iteração, é calculado o valor residual  $r_{im}$ , que é a diferença entre o valor observado e o predito para cada amostra, através da Equação 2.9 (HASTIE et al., 2016).

$$r_{im} = \left[ \frac{\delta L(y_i, F(x_i))}{\delta F(x_i)} \right] \quad (2.9)$$

A Equação 2.9 representa o *Gradient Boosting*, em que é realizada a iteração de  $M$  iterações para  $i = 1, \dots, n$  amostras.

Para cada nó na nova árvore de  $j = 1, \dots, j_m$  são usados os nós finais para gerar um termo de "correção" para a árvore gerada através da Equação 2.10 (HASTIE et al., 2016).

$$\gamma_{jm} = \operatorname{argmin} \sum_{i=1}^n L(y_i, F_m(x_i)) + \gamma \quad (2.10)$$

Na última etapa da iteração a função  $F_m(x)$  é atualizada, e uma nova predição é realizada utilizando a predição anterior  $F_{m-1}(x)$ .

A função  $F_m(x)$  é descrita pela Equação 2.11:

$$F_m(x) = F_{m-1}(x) + v \sum_{i=1}^{j_m} \gamma_{jm} \quad (2.11)$$

onde  $v$  é a taxa de aprendizado, que é o parâmetro que a rapidez com que o modelo aprende a partir dos dados (HASTIE et al., 2016).

Após o final do ciclo de iteração, o valor final da última árvore  $F_m$  finalmente é retornado.

### 3. METODOLOGIA

No LHC, os registros das colisões de partículas resultam em enormes quantidades de dados. No entanto, na maioria dos casos, haverá uma certa quantidade

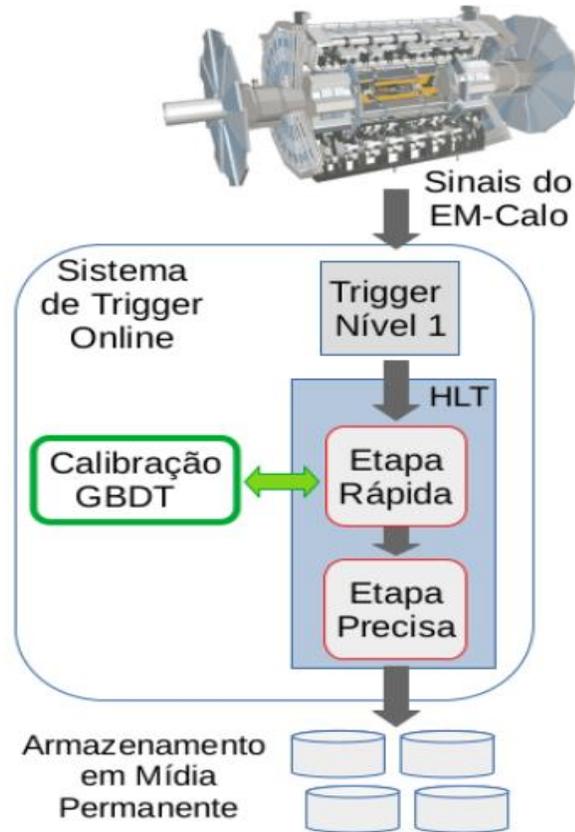
de ruído de fundo contaminando os dados. Diante disso, para que se tenha uma boa classificação dos eventos de interesse, é necessário converter as informações do calorímetro em um valor que represente, de maneira aproximada, a quantidade de energia que a partícula perde ao interagir com o material do sensor (AABOUD et al, 2019).

A energia da partícula eletromagnética é calculada somando as contribuições das células das camadas EM1, EM2 e EM3 localizadas dentro de uma região de interesse (RoI) ao redor da célula de maior energia (célula mais energética). No entanto, é importante destacar que podem ocorrer perdas de informação nas regiões laterais além da área da RoI, nas regiões longitudinais (nas camadas hadrônicas) e devido à interação da partícula com o material do detector antes de atingir o calorímetro (região *upstream*) (SIMAS FILHO, et al., 2021).

Diante disso, o objetivo do procedimento de calibração é corrigir os fatores de erro na medição. Para isso, será incluída uma árvore de decisão (*Gradient Boosted Decision Tree*) na etapa rápida do HLT. A adição dessa etapa de processamento, irá possibilitar a correção a estimativa de energia do agrupamento na entrada do HLT, resultando em um desempenho otimizado da etapa rápida e, uma redução na carga computacional exigida pelo HLT.

A Figura 3.1 demonstra em qual etapa do HLT a árvore de decisão será adicionada. O reforço de uma árvore de decisão indica o "aprendizado" de uma árvore a partir dos candidatos não classificados das árvores de decisão construídas anteriormente.

Figura 3.1 – Diagrama que ilustra a integração do bloco de calibração proposto com o HLT (*High-Level Trigger*) do Experimento ATLAS.



(Fonte: SIMAS FILHO, et al., 2021)

A energia dos elétrons e fótons é calculada a partir da energia do cluster reconstruído, aplicando uma correção para a energia perdida no material a montante do calorímetro, para a energia depositada nas células vizinhas ao aglomerado em  $\eta$  e  $\phi$ , e a energia perdida além do calorímetro (AABOUD et al., 2019).

A energia total depositada em uma célula do calorímetro pode ser estimada pela Equação 3.1:

$$E = \alpha_{ADC \rightarrow MeV} \times \alpha_{BDT} \times \alpha_{MC \rightarrow Data} \times \sum_{j=1}^N a_j (s_j - p) \quad (3.1)$$

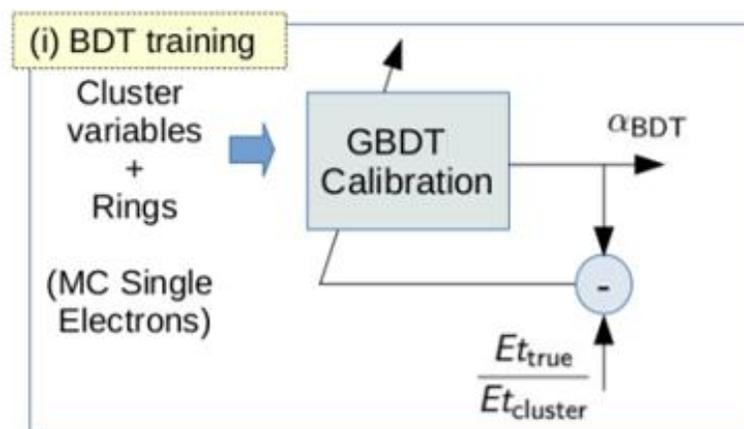
em que:

- $\alpha_{ADC \rightarrow MeV}$  incorpora as conversões do nível de conversão AD para corrente elétrica e, em seguida, de corrente elétrica para MeV. Esse fator é determinado por meio de procedimentos experimentais de calibração (AABOUD, et al., 2019);
- $\alpha_{BDT}$  é o fator de correção fornecido pela GBDT;
- $\alpha_{MC \rightarrow Data}$  trata-se de um fator que realiza o ajuste dos dados simulados (Monte Carlo - MC) aos dados experimentais (Data). Esse fator é necessário, uma vez que a BDT é desenvolvida para os dados simulados (SIMAS FILHO, et al., 2021);
- $a_j$  e  $N$  representam respectivamente os pesos e a ordem do filtro casado são os parâmetros utilizados na detecção dos pulsos médios no calorímetro (AABOUD et al., 2019);
- $s_j$  são as medições obtidas no calorímetro (SIMAS FILHO, et al., 2021);
- $p$  representa o valor médio esperado nos sensores do calorímetro (SIMAS FILHO, et al., 2021).

A etapa realizada para a estimação do fator de conversão  $\alpha_{BDT}$  está representada na Figura 3.2.

O treinamento da GBDT, é realizado usando informações do agrupamento de células e dos anéis como entradas (SIMAS FILHO, et al., 2021).

Figura 3.2 – Diagrama que representa a etapa do sistema de calibração incluído no segundo nível de filtragem do Experimento ATLAS.



(Fonte: SIMAS FILHO, et al., 2021)

A resposta do calorímetro do ATLAS apresenta características distintas de acordo com a energia transversa da partícula,  $E_t$ . Isso ocorre porque o perfil de deposição varia conforme a  $E_t$ , resultando em maior dificuldade na identificação de elétrons em baixas energias (SIMAS FILHO, et al., 2021)

Dessa forma, é necessário projetar todo o procedimento de calibração levando em consideração diferentes intervalos de  $|\eta|$  e  $E_t$ . Neste trabalho foram adotados os seguintes intervalos:

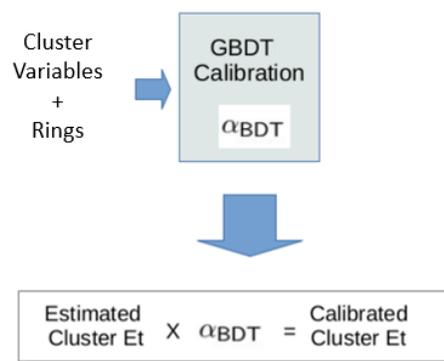
- $|\eta| = [0, 0.6, 0.8, 1.2, 1.37, 1.52, 1.8, 2.0, 2.2, 2.5]$
- $E_t = [0, 5, 10, 15, 20, 30, 40, 50, 70, 100, 150, 200, 250, 900, 3000] \text{ GeV}$ .

Os fatores  $\alpha$  descritos na **Equação 3.1** devem ser definidos para cada par de combinação  $|\eta| \times E_t$ , resultando em um conjunto de 117 fatores distintos.

Neste estudo, foi empregado um conjunto de árvores de decisão com reforço por gradiente e para o treinamento, utilizou-se o framework *LightGBM*.

A principal característica do *LightGBM* é o uso de uma estratégia de crescimento de árvores chamada de *leaf-wise*. Nesse contexto, o *LightGBM* constrói as árvores de escolhendo os nós que fornecem a maior redução no valor da função de perda durante o treinamento. Isso resulta em uma construção mais rápida das árvores e, potencialmente, em um melhor desempenho preditivo (KE, G. et al., 2017).

Figura 3.3 – Representação das etapas de operação do sistema de calibração proposto.



(Fonte: Adaptado SIMAS FILHO, et al., 2021)

A calibração da GBDT utiliza como entradas variáveis descritoras do formato do perfil de deposição nos calorímetros. Um total de 6 atributos derivados do perfil de deposição no calorímetro são utilizados como entradas para as árvores de decisão (SIMAS FILHO, et al., 2021).

A saída da GBDT é o fator de conversão  $\alpha_{BDT}$  como mostra a Equação 3.2

$$\alpha_{BDT} = \frac{E_{truth}}{E_{fast}} \quad (3.2)$$

A GBDT utiliza os anéis concêntricos de deposição de energia como entrada, que representam a distribuição de energia tanto longitudinal quanto transversalmente. No entanto, estudos iniciais indicaram a presença de assimetrias na deposição de energia em relação aos eixos espaciais de referência  $\eta$  e  $\varphi$  do calorímetro (SIMAS FILHO, et al., 2021).

Para identificar a influência dos anéis concêntricos na calibração foram realizados treinamentos utilizando 28 anéis e 100 anéis, a fim de identificar a diferença de desempenho da calibração nos dois cenários.

Os 28 anéis utilizados foram selecionados para manter a maior parte da informação relevante para o problema de calibração. Deste modo, foram privilegiados os anéis centrais das camadas que corrigem para a energia perdida antes (PS) e depois (HAD1) do calorímetro e também da camada que amostra maior quantidade de energia eletromagnética (EM2). Neste contexto foram selecionados os anéis indicados na Tabela 3.1

Tabela 3.1 - Tabela com as quantidades de anéis selecionados para cada camada.

PS	EM1	EM2	EM3	HAD1	HAD2	HAD3
4	8	8	4	4	0	0

Os treinamentos foram realizados para os dois casos sob as mesmas condições com o intuito avaliar as diferenças existentes entre as calibrações, e o efeito delas na classificação dos eventos de interesse.

Para analisar o impacto do uso dessas configurações de entradas no treinamento da GBDT foram construídos gráficos de erro relativo, distância interquartílica e *scatter plot* que comparam os resultados entre os dois arranjos.

A distância interquartílica (do inglês: *Interquatile Range (IQR)*) é uma medida estatística que descreve a dispersão ou variabilidade dos dados em um conjunto de observações (NAVIDI, 2006). Sendo assim, ao comparar diferentes distribuições de dados utilizando a IQR pode-se avaliar a variabilidade relativa entre elas, uma vez que, uma distribuição com um IQR maior geralmente indica maior dispersão e variabilidade dos dados.

Uma outra forma de comparar o impacto na calibração do treinamento da GBDT com o conjunto completo de anéis e o reduzido é através do histograma do erro relativo e do *scatter plot*. O primeiro é uma representação gráfica que permite visualizar a distribuição dos erros relativos de um, conjunto de dados. Dessa forma, pode-se visualizar se existem tendências, concentrações ou dispersões específicas nos erros. O segundo, auxilia na comparação entre os resultados das duas configurações e identificar possíveis relações ou padrões entre elas. Para o cenário desse trabalho, foi analisado a relação  $x=y$ , ou seja, se os resultados obtidos com a GBDT treinada com a entrada de 28 anéis é similar aos da treinada com os 100 anéis.

## 4. RESULTADOS

Para analisar os resultados da calibração e comparar espalhamento da razão  $E_{truth}/E_{fast}$ , foi calculada a distância interquartilica, isso porque ela indica a região em torno da mediana em que se concentra 50% da área da PDF (*Probability Density Function*).

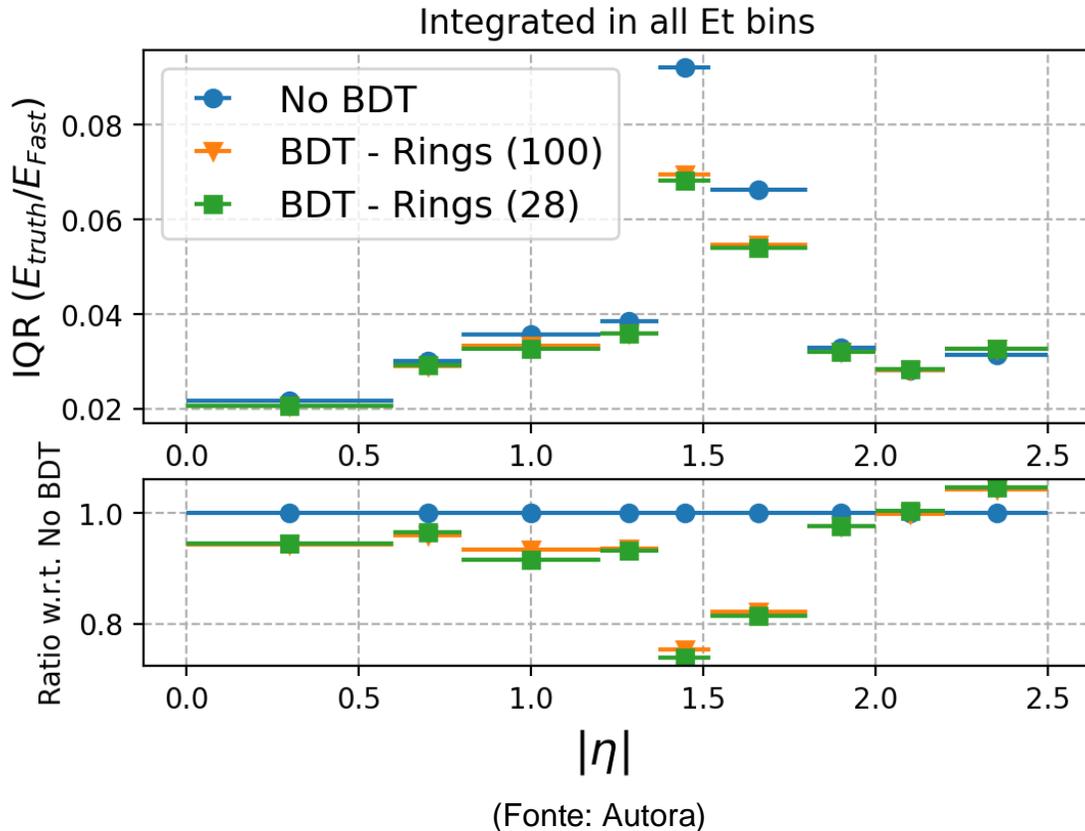
Para a implementação da GBTDE, foi empregado um conjunto de dados composto por simulações de Monte Carlo de colisões que geram elétrons isolados. Cerca de 500.000 eventos foram utilizados, com distribuição de energia variando de 0 a aproximadamente 2500 GeV.

A GBTDE foi treinada considerando dois cenários:

- *Rings28*: treinamento realizado considerando como entrada apenas vinte e oito anéis concêntricos.
- *Rings100*: treinamento realizado considerando como entrada todos anéis.

Após o treinamento, foram construídos gráficos, representados pelas Figura 4.1 e Figura 4.2, para analisar o comportamento da IQR (*Interquartile Range*) normalizada pela PDF da calibração padrão (sem anéis) para diferentes intervalos de  $E_t$  e  $\eta$ . Essa análise é importante, porque através dela pode-se obter informações sobre a distribuição e a variabilidade dos dados.

Figura 4.1 – Gráfico comparativo da IQR para diferentes intervalos de  $\eta$  (pseudorapidez) para a calibração padrão, e para a BDT treinada com 28 e 100 anéis.

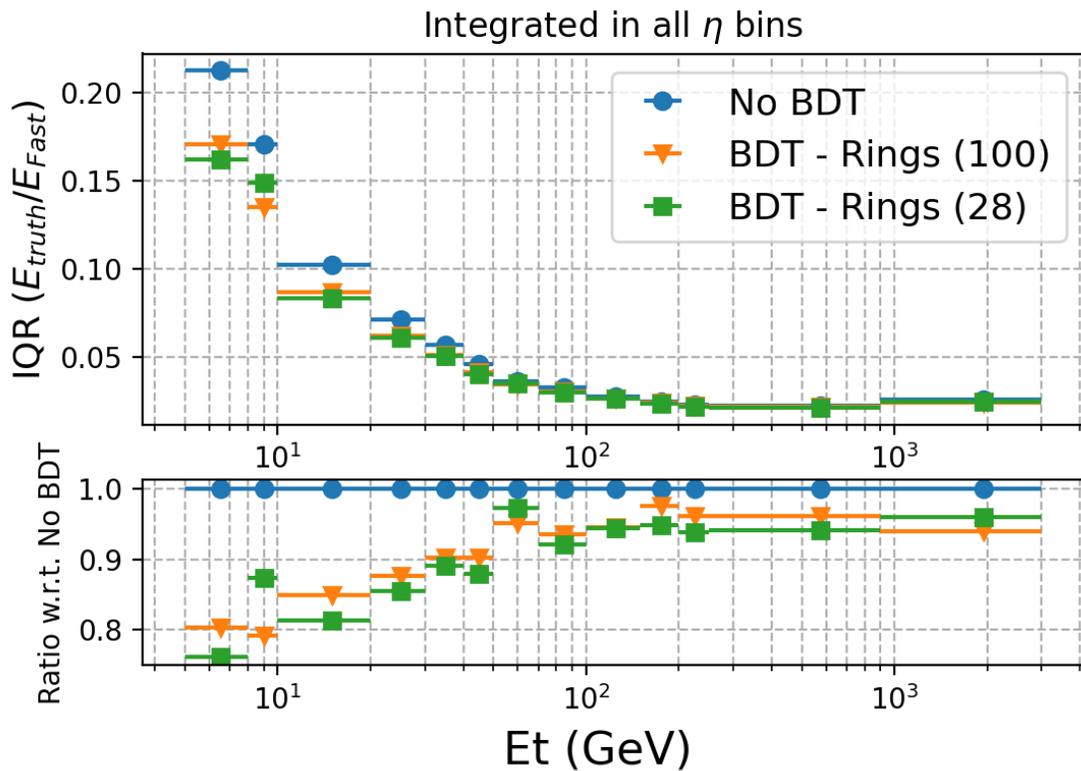


A Figura 4.1 mostra que a dispersão foi relativamente parecida para os dois casos de calibração com a BDT. Além disso, para os valores de  $\eta$  próximos a 1.5 a dispersão da calibração utilizando a BDT foi muito menor, em relação a calibração padrão.

Comparado os cenários de treinamento da BDT *Rings100* e *Rings28* é possível perceber que eles apresentam comportamentos muito similares. O gráfico inferior representa a variação relativa da BDT calibrada com os 100 e 28 anéis, e calibração padrão (*noBDT*). De maneira geral o comportamento as duas BDT calibrada com os 100 anéis e 28 anéis são similares, e no comparativo com a calibração sem BDT, apresentam resultado com menor dispersão conforme medido pelo IQR.

A mesma análise de IQR foi realizada para os intervalos de  $E_t$  (energia transversa) e é descrita na Figura 4.2.

Figura 4.2 – Gráfico comparativo da IQR para diferentes intervalos de  $E_t$  para a calibração padrão, e para a BDT treinada com 28 e 100 anéis.



(Fonte: Autora)

Para os intervalos de  $E_t$  a Figura 4.2 demonstra que a dispersão foi relativamente parecida para os dois casos de calibração com a BDT e para o caso de calibração padrão (sem a BDT). O gráfico inferior representa a variação relativa da BDT calibrada com os 100 anéis e com os 28 anéis a calibração padrão (sem a BDT). Através dele, pode-se perceber que a calibração com as duas BDT calibrada com os 100 anéis e 28 anéis se comportam de maneira similar, e no comparativo com a calibração sem BDT, também apresentam resultados de dispersão melhores.

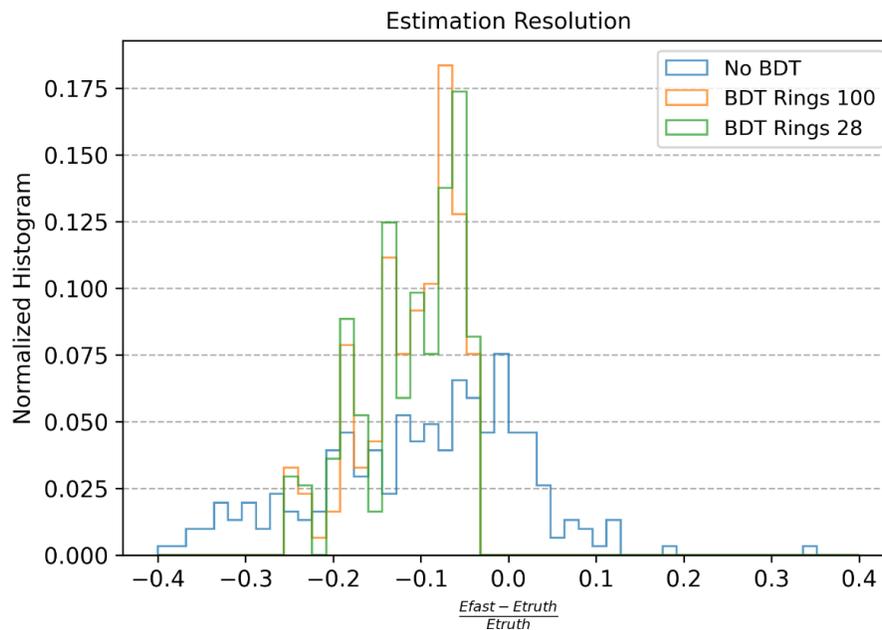
Tendo em vista da análise de modelos preditivos, a relação entre calibração e dispersão indica à qualidade do ajuste do modelo em relação aos dados observados.

A calibração de um modelo preditivo envolve ajustar os parâmetros do modelo de forma que ele seja capaz de prever os valores observados o mais precisamente possível. Diante disso, a análise da dispersão nesses cenários permite conhecer a variabilidade entre as previsões feitas pelo modelo e os valores reais correspondentes. Uma dispersão alta indica que as previsões do modelo estão mais espalhadas e menos precisas em relação aos dados observados, enquanto uma dispersão baixa indica que as previsões são mais precisas e consistentes.

Uma outra forma de avaliar os resultados das calibrações é comparar os erros relativos para intervalos  $E_t$ . O erro relativo é calculado como a diferença entre o valor real ( $E_{truth}$ ) e o valor previsto pela etapa rápida ( $E_{fast}$ ), dividida pelo valor real.

As Figuras 4.3 e 4.4 representam os histogramas do erro relativo para intervalos de  $E_t$  de 5 a 8 GeV, sendo a segunda em escala logarítmica.

Figura 4.3 – Histograma do erro relativo das energias estimadas sem a BDT e com a BDT com as configurações de entrada *Rings28* e *Rings100* – intervalo de [5,8] GeV.

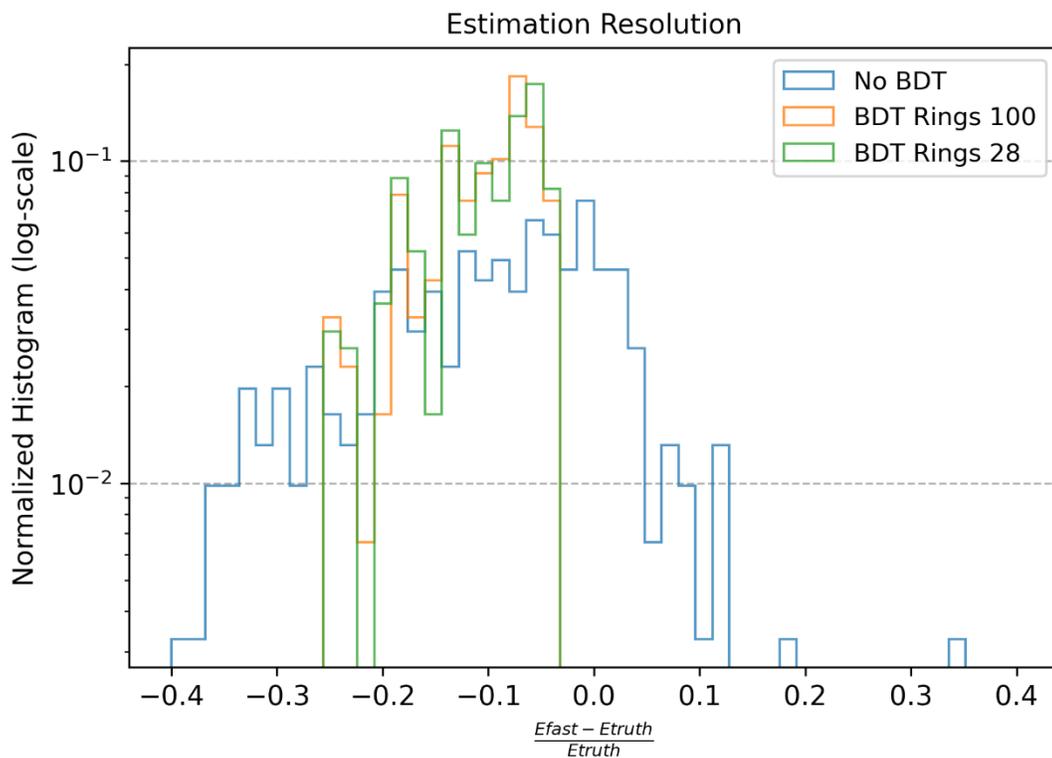


(Fonte: Autora)

A Figura 4.3 demonstra que os eventos estimados pela calibração com a BDT *Rings28* e *Rings100* estão mais concentrados onde o erro é zero, diferente do que acontece para os eventos estimados pela calibração sem a BDT.

Apesar do pico expressivo quando o erro é zero, nos intervalos de baixas energias a precisão das estimativas não é tão boa uma vez que muitos eventos foram estimados para o erro próximo de 20% para o caso da calibração com a BDT, e acima de 20% para os estimados sem a BDT. Essa diferença pode ser melhor visualizada na escala logarítmica como mostra a Figura 4.4.

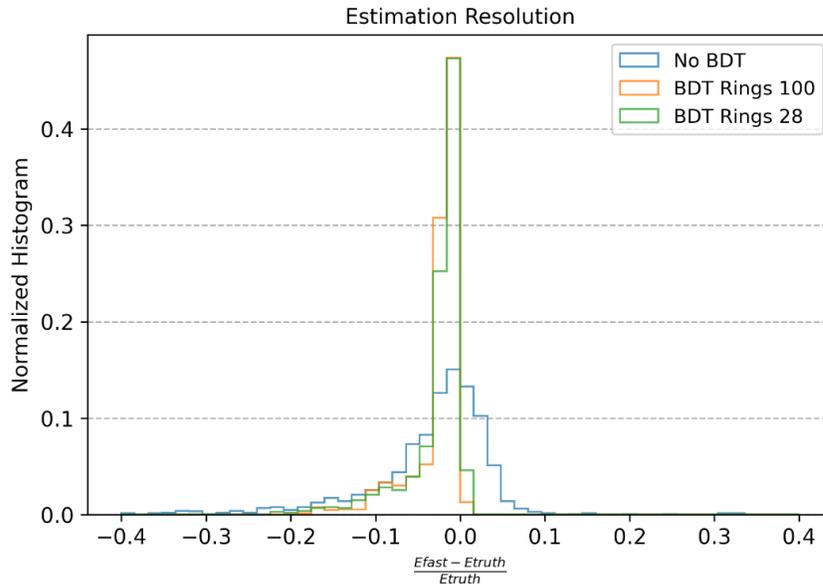
Figura 4.4 – Histograma do erro relativo das energias estimadas sem a BDT e com a BDT com as configurações de entrada *Rings28* e *Rings100* em escala logarítmica – intervalo de [5,8] GeV.



(Fonte: Autora)

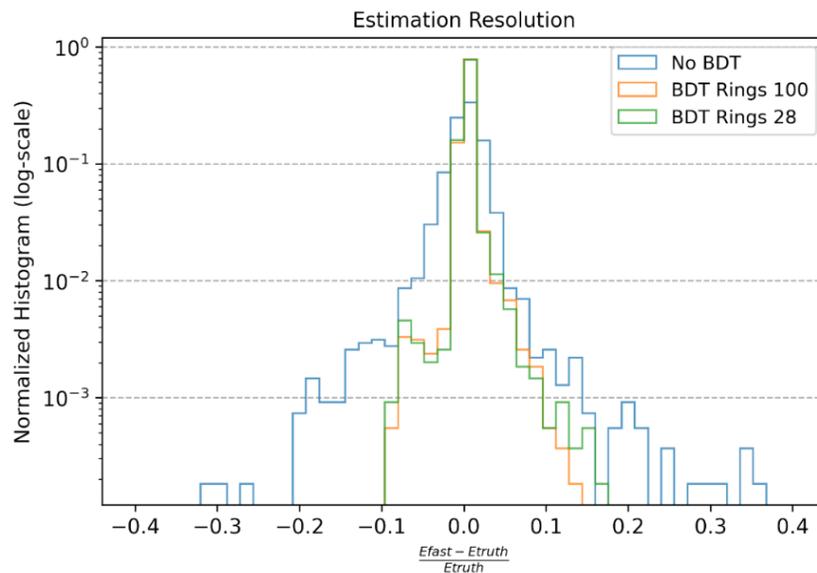
A mesma análise foi feita para outros intervalos de energia e os resultados obtidos são descritos nas Figuras 4.5, 4.6, 4.7.

Figura 4.5 – Histograma do erro relativo das energias estimadas sem a BDT e com a BDT com as configurações de entrada *Rings28* e *Rings100* – intervalo de [20,30] GeV.



(Fonte: Autora)

Figura 4.6 – Histograma do erro relativo das energias estimadas sem a BDT e com a BDT com as configurações de entrada *Rings28* e *Rings100* em escala logarítmica – intervalo de [20,30] GeV.



(Fonte: Autora)

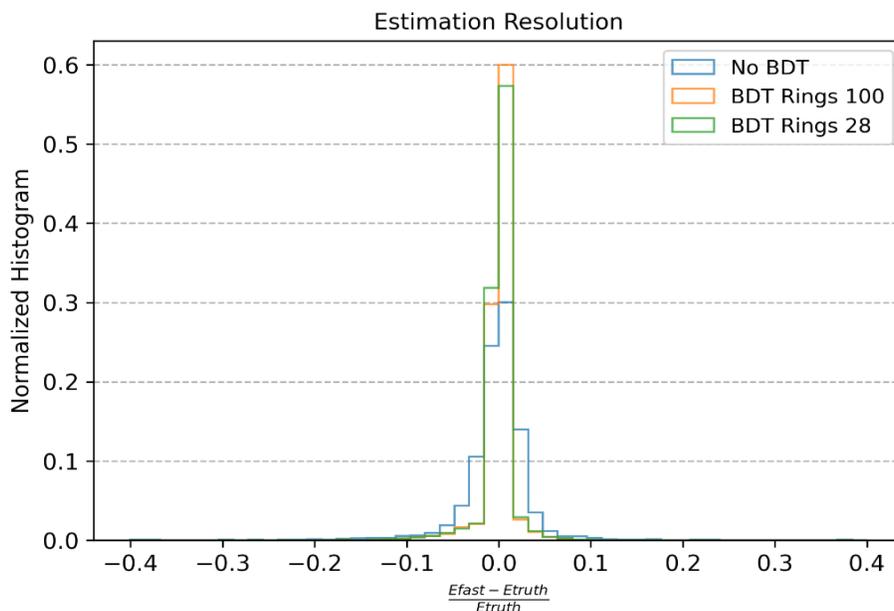
Para o intervalo de [20,30] a maioria dos erros foi concentrada em torno de zero para os casos de calibração com BDT, como mostra a Figura 4.5, e houve uma diminuição de estimação de eventos onde é maior que 20%. No entanto, no caso da calibração sem a BDT, o pico quando o erro é zero foi aproximadamente 30% menor que os da calibração com a BDT.

Para o caso da calibração sem a BDT muitos eventos foram estimados para o erro maior que 20% como mostra o histograma na escala logarítmica da Figura 4.6.

Foi construído o histograma de erro relativo para intervalos de energia maiores, nesses casos como mostram as Figuras 4.7 4.8 4.9, 4.10 as calibrações com a BDT tiveram uma distribuição muito mais concentrada quando comparada com a calibração sem a BDT.

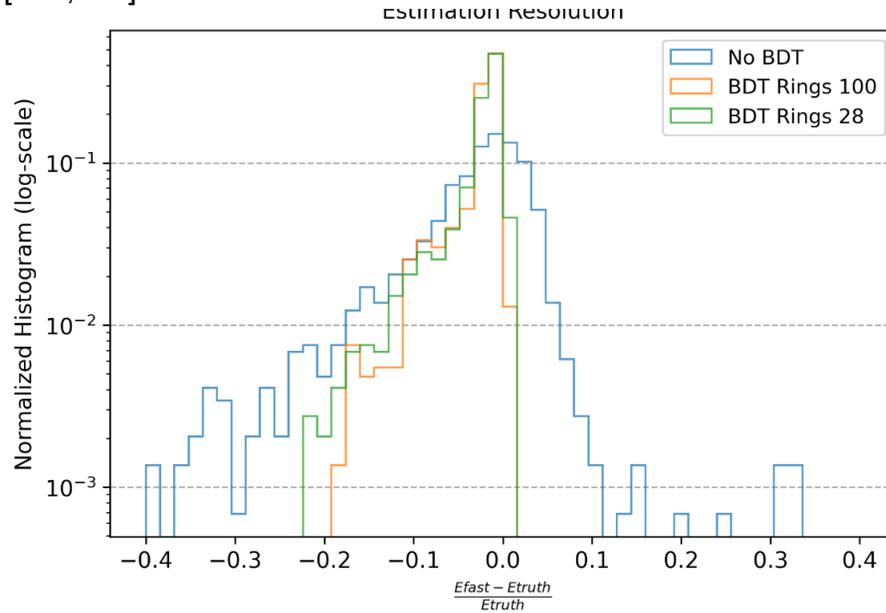
Como aconteceu nos intervalos de baixa energia, a calibração sem a BDT apresenta uma distribuição mais espalhada e menos concentrada em torno do zero, como mostram os histogramas na escala logarítmica as Figuras 4.8, 4.10.

Figura 4.7 – Histograma do erro relativo das energias estimadas sem a BDT e com a BDT com as configurações de entrada *Rings28* e *Rings100* – intervalo de [100,150] GeV.



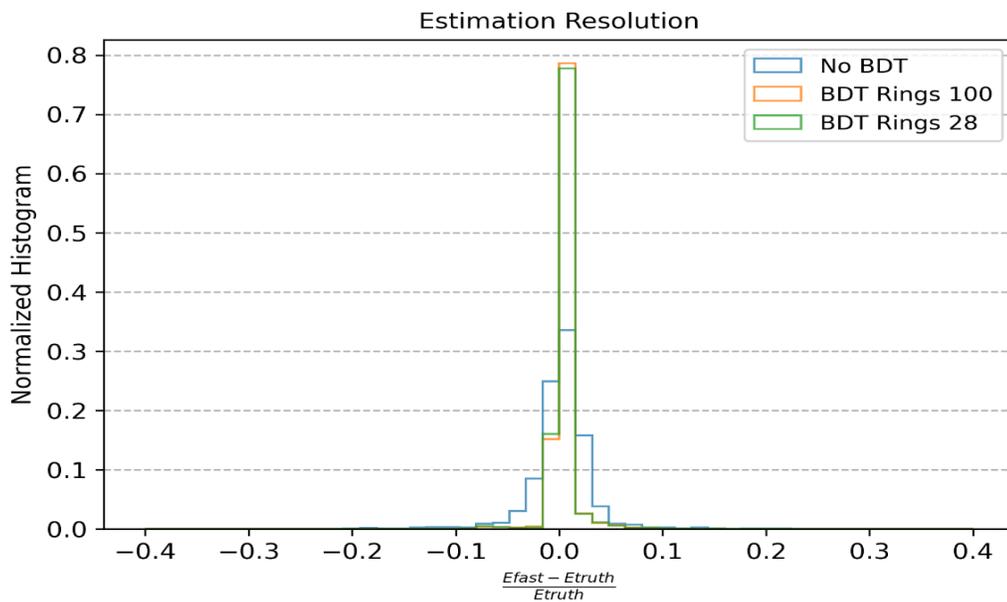
(Fonte: Autora)

Figura 4.8 – Histograma do erro relativo das energias estimadas sem a BDT e com a BDT com as configurações de entrada *Rings28* e *Rings100* em escala logarítmica – intervalo de [100,150] GeV.



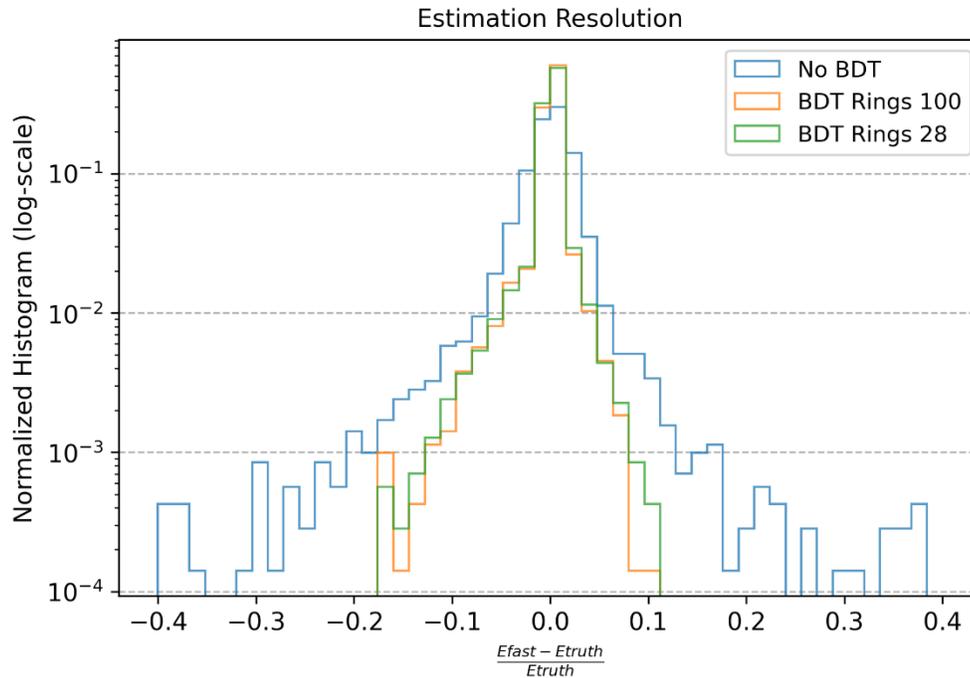
(Fonte: Autora)

Figura 4.9 – Histograma do erro relativo das energias estimadas sem a BDT e com a BDT com as configurações de entrada *Rings28* e *Rings100*.escrever – intervalo de [150,200] GeV.



(Fonte: Autora)

Figura 4.10 – Histograma do erro relativo das energias estimadas sem a BDT e com a BDT de [150,200] GeV.



No contexto de calibração, um histograma do erro relativo pode ser usado para analisar a precisão e a qualidade do processo de calibração. Quando a maioria dos erros se concentra em torno de zero, isso pode indicar que a calibração foi bem-sucedida e precisa, como aconteceu para os casos da calibração com a BDT nos intervalos de alta energia. Quando acontece uma assimetria, isso pode indicar possíveis erros sistemáticos ou imprecisões na calibração, o que foi observado na calibração sem a BDT.

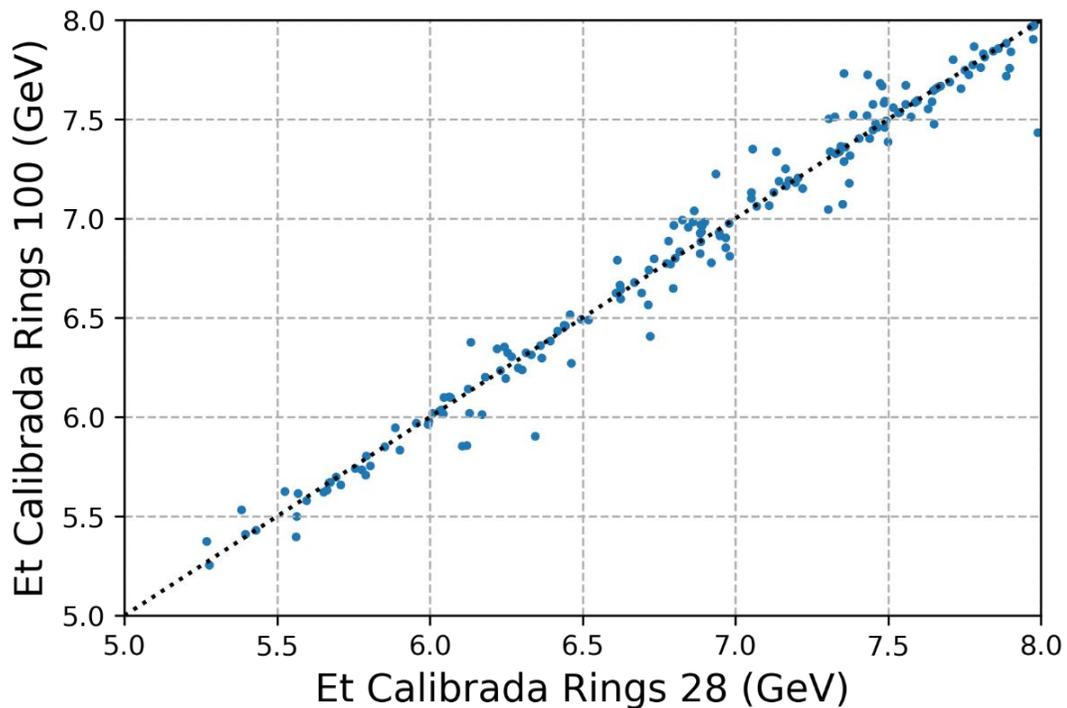
De modo geral a distribuição dos erros para a calibração com a BDT utilizando a configuração de *Rings28* e *Rings100* foram similares, no entanto é possível perceber que a dispersão dos erros para o caso *Rings100* é menor que para o caso *Rings28*. Comportamento este que pode ser justificado pelo fato de que ao utilizar todos os anéis como entrada, mais eventos são utilizados no treinamento da árvore de decisão, o que em contrapartida onera o tempo de treinamento demanda mais recurso computacional.

Para analisar e comparar comportamento da calibração da GBT usando os cem anéis e os vinte e oito anéis, também foi gerado um gráfico de dispersão do tipo *scatter plot* para quatro intervalos de energia. Os gráficos foram construídos, para os mesmos intervalos de energia usados nas análises dos histogramas de erro relativo.

Nesse gráfico, cada ponto representa um par de valores correspondentes às duas variáveis em estudo. Através dele é possível analisar se os pontos estão fortemente agrupados em torno da primeira bissetriz ( $x=y$ ), nos intervalos de energia selecionados.

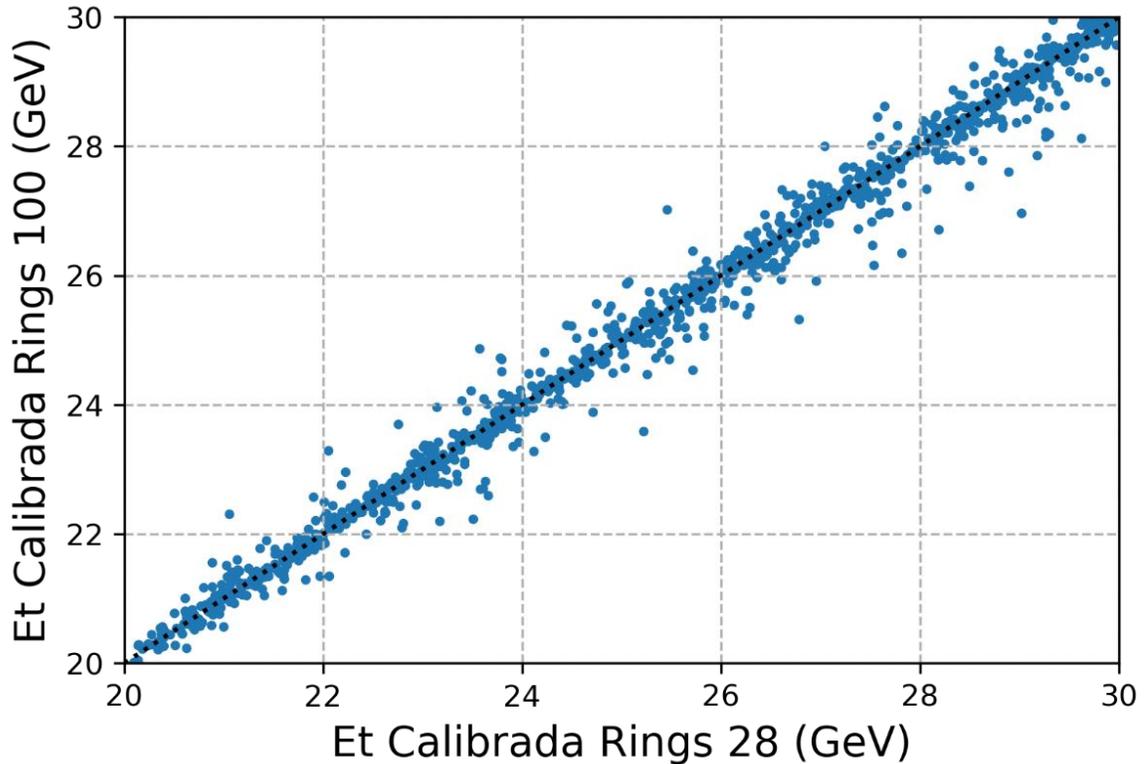
Para os intervalos de baixa energia, como mostram as Figuras 4.11 e 4.12, é possível perceber que a uma dispersão maior dos pontos, e isso pode ser consequência da dispersão maior dos erros de calibração para intervalos menores de energia.

Figura 4.11 – *Scatter plot* da calibração com a BDT com as configurações de entrada *Rings28* e *Rings100* – intervalo de [5,8] GeV.



(Fonte: Autora)

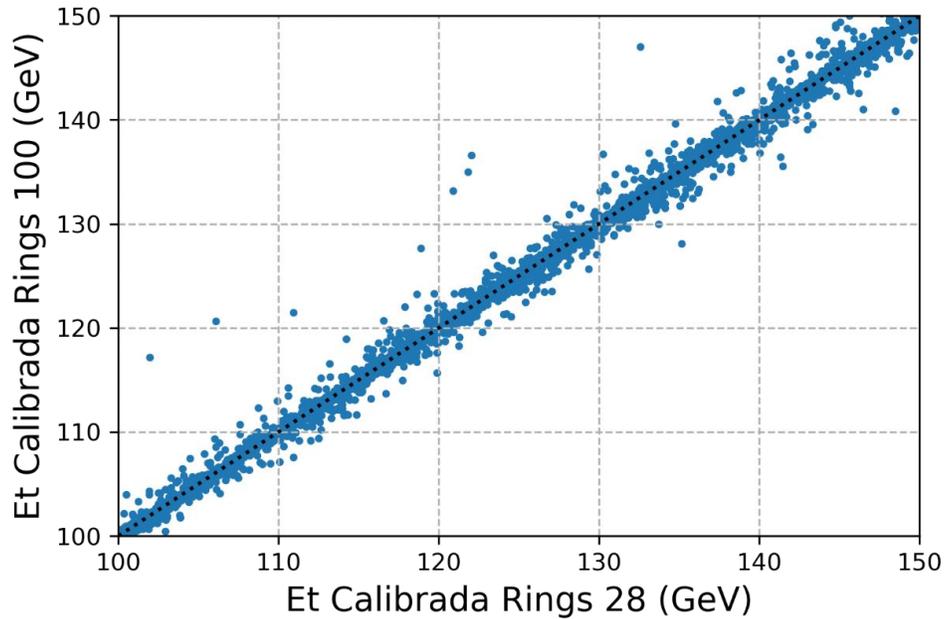
Figura 4.12 – Scatter plot da calibração com a BDT com as configurações de entrada Rings28 e Rings100 – intervalo de [20,30] GeV.



(Fonte: Autora)

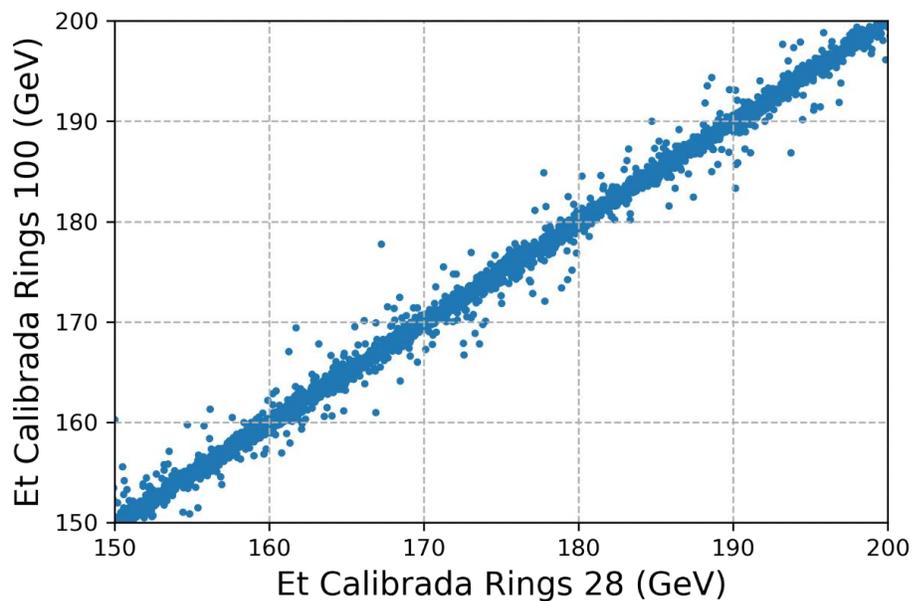
Para os intervalos de alta energia, como mostram as Figuras 4.13 e 4.14, é possível perceber que a uma concentração maior em torno da primeira bissetriz. Isso mostra que, as estimações dos eventos para as calibrações utilizando a BTD com a configuração de entrada *Rings100* e *Rings28*, no geral se comportaram de maneira semelhante, uma vez que uma dispersão menor indica que as previsões dos dois modelos são consistentes entre si.

Figura 4.13 – *Scatter plot* da calibração com a BDT com as configurações de entrada *Rings28* e *Rings100* – intervalo de [100,150] GeV.



(Fonte: Autora)

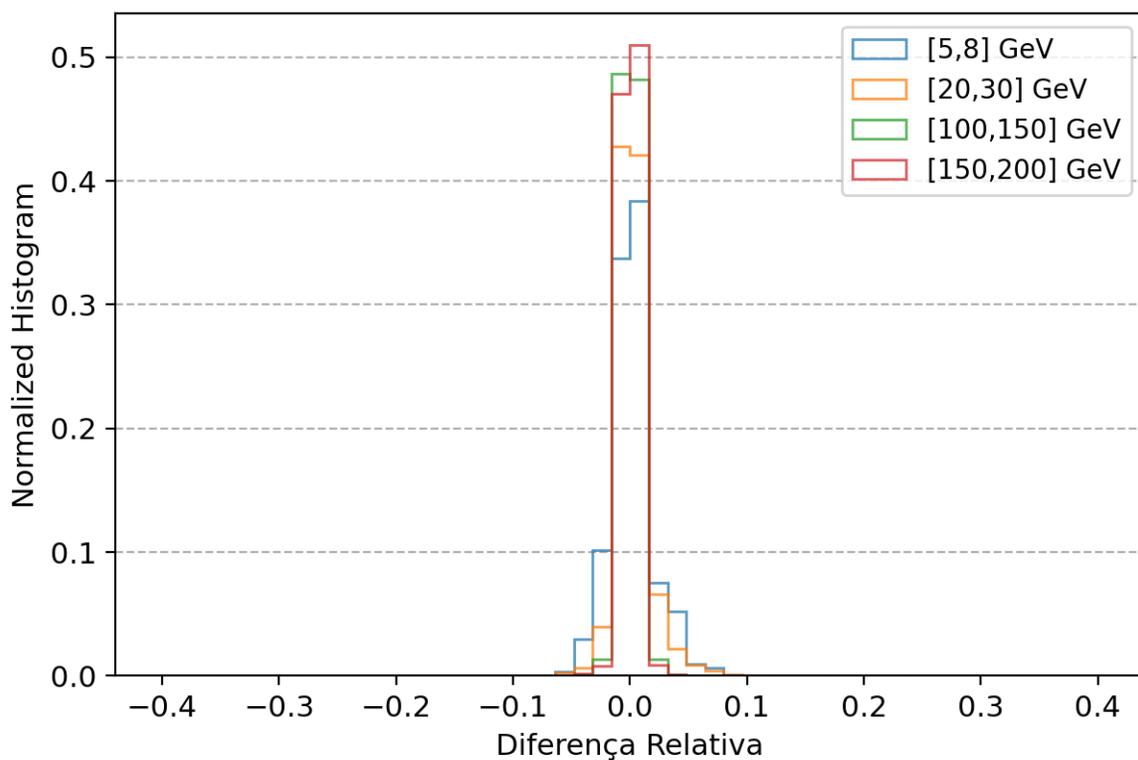
Figura 4.14 – *Scatter plot* da calibração com a BDT com as configurações de entrada *Rings28* e *Rings100* – intervalo de [150,200] GeV.



(Fonte: Autora)

Além do *scatter plot*, também foi construído o histograma da diferença relativa para quatro intervalos de energia diferentes, como mostra a Figura 4.15. A diferença relativa foi calculada como a diferença entre os valores de energia estimados pela calibração utilizando o conjunto completo de anéis e o valor estimado usando o conjunto reduzido.

Figura 4.15 – Histograma da diferença relativa das energias estimadas com a BDT com as configurações de entrada Rings28 e Rings100



(Fonte: Autora)

A Figura 4.15 indica que a diferença relativa entre as BDTs com treinadas com os 28 e 100 anéis são muito próximas a zero, demonstrando grande similaridade entre os cenários.

## 4.1 Análise dos Resultados

Através dos resultados obtidos da implementação da calibração na etapa rápida do HLT utilizando *Gradient Boosted Decision Tree*, observou-se que os resultados obtidos para o treinamento utilizando o conjunto menor de anéis, de modo geral, é similar ao do de configuração com cem anéis.

A Figuras 4.1 e 4.2 mostraram que para as duas configurações de calibração da GBDT, os resultados de dispersão das estimativas dos eventos foram melhores que sem a calibração, além de serem similares. Dessa maneira, o treinamento com Rings28 apresenta um resultado adequado para ser utilizado na etapa rápida do HLT.

Os Histogramas do erro relativo mostraram que para intervalos mais baixos de energia, como mostram as Figuras 4.3 e 4.4, muitos eventos foram estimados para o erro próximo de 20%, para ambas as configurações de entrada da GBDT (Rings28 e Rings100), No entanto, a maioria dos erros se concentraram em torno de zero, para os casos da calibração com a BDT nos intervalos de alta energia (Figuras 4.7 e 4.9).

De modo geral as estimações dos eventos para as calibrações utilizando a BDT com a configuração de entrada completa e reduzida dos anéis, tiveram resultados parecidos, como mostram as Figuras 4.13,4.14., No entanto, para os intervalos de baixa energia, como mostram as Figuras 4.11 e 4.12, houve um espalhamento maior dos pontos, e isso pode ser consequência da dispersão maior dos erros de calibração para intervalos menores de energia.

A análise dos resultados obtidos demonstra que o treinamento da GBDT com a configuração de cem anéis e vinte e oito anéis apresentam resultados parecidos, como observado na Figura 4.15. Diante disso, afim de diminuir a demanda computacional e tempo de treinamento, o conjunto de anéis menor pode ser usado na calibração da etapa do HLT para melhorar a estimacão dos eventos.

## 5. CONCLUSÃO

Neste trabalho, foi apresentada uma nova metodologia para a calibração das medições de energia de elétrons na etapa rápida de trigger do Experimento ATLAS, e foi estudado o comportamento da mesma em dois cenários.

Durante o desenvolvimento deste trabalho, foi realizada uma revisão da literatura relacionada ao sistema de filtragem online do Experimento ATLAS e da calibração por árvore de decisão e à sua aplicação na etapa rápida do HLT. Isso proporcionou uma compreensão dos fundamentos teóricos e dos métodos envolvidos, bem como das suas vantagens e desafios.

A etapa rápida do HLT deve aplicar critérios mais detalhados do que o nível de filtragem anterior (L1 Trigger). Isso permite a identificação de assinaturas de partículas raras ou eventos de interesse em meio a uma elevada quantidade de eventos conhecidos e não interessantes. Por isso, a filtragem eficiente, nessa fase, ajuda a otimizar o uso dos recursos computacionais disponíveis.

A implementação do sistema de calibração por árvore de decisão no HLT do Experimento ATLAS revelou resultados promissores. Os treinamentos realizados demonstraram que o treinamento da GBTD com a configuração de 28 anéis apresenta resultados similares ao do treinamento com 100 anéis. Dessa forma ao implementar esse modelo de calibração no HLT pode-se obter melhorias significativas na reconstrução dos eventos com uma baixa demanda computacional.

### 5.1. Trabalhos Futuros

Com base no estudo realizado sobre a etapa rápida do High-Level Trigger (HLT) do Experimento ATLAS, foram identificadas algumas possíveis direções para trabalhos futuros que visam aprimorar ainda mais essa importante etapa do experimento.

Um estudo interessante consiste na otimização dos algoritmos de seleção de eventos utilizados no HLT. É fundamental explorar métodos avançados de filtragem e

critérios de decisão para melhorar a eficiência de seleção e reduzir as taxas de eventos falsos. Para isso podem ser considerados o uso de outras técnicas de aprendizado de máquina em substituição a GBDT que podem proporcionar avanços significativos no desempenho do HLT.

Além disso, formas otimizadas de escolher os anéis mais relevantes, utilizando por exemplo a informação mútua entre cada entrada e a saída da predição podem ser estudadas, bem como a realização de testes de calibração no sistema computacional do Experimento ATLAS.

## REFERÊNCIAS

- AABOUD, M et al. Electron and photon energy calibration with the ATLAS detector using 2015-2016 LHC proton-proton collision data. **Journal of instrumentation**, 2019.
- ACHENBACH, R. et al. The ATLAS level-1 calorimeter trigger. **Journal of Instrumentation**, v. 3, n. 03, p. P03001, 2008.
- ALPAYDIN, E. **Introduction to machine learning**. MIT press, 2014.
- ANJOS, A. dos et al. Neural triggering system operating on high resolution calorimetry information. *Nuclear Instruments and Methods In Physics Research A*, v. 559, p. 134–138, 2006.
- ANJOS, A. Sistema online de Filtragem em um Ambiente com alta taxa de eventos. Tese de Doutorado, COPPE/UFRJ, 2006.
- CERN. LHC guide. 2017. Disponível em: <<https://cds.cern.ch/record/2255762>>.
- CERN. The accelerator complex. 2023. Disponível em: <<https://home.web.cern.ch/about>>. Acesso em Abril de 2023.
- CERN. The Large Hadron Collider. 2023 Disponível em: <<https://home.web.cern.ch/about>>. Acesso em Abril de 2023.
- CIODARO, T. et al. Online particle detection with neural networks based on topological calorimetry information. In: **Journal of physics: conference series**. IOP Publishing, 2012. p. 012030.
- COLLABORATION, ATLAS et al. The ATLAS experiment at the CERN large hadron collider. 2008.
- COSTA, P. R. O impacto da física na medicina moderna. **Jornal da USP**, 2018.
- EVANS, L.; BRYANT, P. LHC machine. **Journal of instrumentation**, v. 3, n. 08, p. S08001, 2008.
- GRIFFITHS, D. **Introduction to elementary particles**. John Wiley & Sons, 2020.

HASTIE, T; TIBSHIRANI, R; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. New York: springer, 2016.

HAYKIN, S. **Neural networks and learning machines, 3/E**. Pearson Education India, 2009.

KE, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. **Advances in neural information processing systems**, v. 30, 2017.

KUHN, M; JOHNSONL, K. **Applied predictive modeling**. New York: Springer, 2013.

MARTINS, Marcos Nogueira; SILVA, Tiago Fiorini da. Electron accelerators: History, applications, and perspectives. **Radiation Physics and Chemistry**, v. 95, p. 78-85, 2014.

MARTINS, M. N.; SILVA, T. F. Electron accelerators: History, applications, and perspectives. **Radiation Physics and Chemistry**, v. 95, p. 78–85, 2014.

MOREIRA, M., A. O modelo padrão da física de partículas. **Revista Brasileira de Ensino de Física**, 2009.

MURPHY, K., P. **Machine learning: a probabilistic perspective**. MIT press, 2012.

NAVIDI, W. C. **Statistics for engineers and scientists**. New York: McGraw-Hill, 2006.

NESSI, M. et al. **ATLAS high-level trigger, data-acquisition and controls: Technical Design Report**. ATLAS-TDR-016, 2003.

PERANTONI, M. **CALIBRAÇÃO E FILTRAGEM COM CALORIMETRIA DE ALTAS ENERGIAS**. 2012. Tese de Doutorado. Universidade Federal do Rio de Janeiro.

RUSSELL, S; NORVIG, P. Artificial intelligence: a modern approach, global edition 4th. **Foundations**, 2020.

SILVA, T., F. Electron accelerators: History, applications, and perspectives. **Radiation Physics and Chemistry**, 2014.

SIMAS FILHO, E. F. de. Análise Não-Linear de Componentes Independentes para uma Filtragem Online Baseada em Calorimetria de Alta Energia e com Fina Segmentação. 310 p. Tese de Doutorado — UFRJ, 2010.

SIMAS FILHO, et al. **Uma proposta para Calibração do Sistema Online de Seleção de Eventos no Detector ATLAS utilizando *Gradient Boosted Decision Trees***, 2014.

SOUZA, E. E. P. **Pré-Processamento Eficiente para um Sistema online de Classificação Baseado em Redes Neurais Artificiais**. 2015. Tese de Doutorado. Dissertação (Mestrado)—UFBA, Salvador, 2015. Disponível em:< <http://www.ppgee.eng.ufba.br/teses/2563ead70510e2565a35c28e9b4377fd.pdf>>. Acesso em Abril de 2023.

THOMSON, M. **Modern particle physics**. Cambridge University Press, 2013.

TIPLER, P. A.; LIEWELLYN, R. A. **Física Moderna**, Rio de Janeiro, 2017.

TORRES, R. Sistema online de filtragem em um ambiente com alta taxa de eventos e fina granularidade. **Rio de Janeiro, UFRJ/COPPE**, 2010.

WIGMANS, R. **Calorimetry: Energy measurement in particle physics**. Oxford University Press, 2000.