UNIVERSIDADE FEDERAL DA BAHIA

ESCOLA POLITÉCNICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIAS ELÉTRICA E DE COMPUTAÇÃO

ARTHUR SODRÉ DE OLIVEIRA ALVES

Dissertação de Mestrado

CALIBRAÇÃO DE ENERGIA NA ETAPA RÁPIDA DO TRIGGER DE ELÉTRONS DO EXPERIMENTO ATLAS UTILIZANDO ÁRVORES DE DECISÃO REFORÇADAS POR GRADIENTE

Salvador

23 de Julho de 2025

Ficha catalográfica elaborada pela Biblioteca Bernadete
Sinay Neves, Escola Politécnica – UFBA.

A474 Alves, Arthur Sodré de Oliveira.
Calibração de energia na etapa rápida trigger de elétrons do experimento atlas utilizando árvores de decisão reforçadas por gradiente / Arthur Sodré de Oliveira Alves. – Salvador, 2025.
79f.: il.

Orientador: Prof. Dr. Eduardo F. de Simas Filho.
Coorientador: Prof. Dr. Juan Lieber Marin.

Dissertação (mestrado) – Programa de Pós-Graduação em Engenharias Elétrica e de Computação, Escola Politécnica, Universidade Federal da Bahia, 2025.

1. Física – alta energia. 2. Atlas. 3. Energia - calibração. 4. Aprendizado por máquina. I. Simas Filho, Eduardo F. de. II. Lieber Marin, Juan. III. Universidade Federal da Bahia. IV. Título.

CDD: 530

Arthur Sodré de Oliveira Alves

CALIBRAÇÃO DE ENERGIA NA ETAPA RÁPIDA DO TRIGGER DE ELÉTRONS DO EXPERIMENTO ATLAS UTILIZANDO ÁRVORES DE DECISÃO REFORÇADAS POR GRADIENTE

Dissertação de Mestrado

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharias Elétrica e de Computação da Universidade Federal da Bahia como um dos requisitos para obtenção do grau de Mestre em Engenharias Elétrica e de Computação.

Orientador: Prof. Dr. Eduardo F. de Simas Filho
Coorientador: Prof Dr. Juan Lieber Marin

Salvador
23 de Julho de 2025

Arthur Sodré de Oliveira Alves

CALIBRAÇÃO DE ENERGIA NA ETAPA RÁPIDA DO TRIGGER DE ELÉTRONS DO EXPERIMENTO ATLAS UTILIZANDO ÁRVORES DE DECISÃO REFORÇADAS POR GRADIENTE

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharias Elétrica e de Computação da Universidade Federal da Bahia como parte dos requisitos para obtenção do grau de Mestre em Engenharias Elétrica e de Computação.

Trabalho aprovado. Salvador, 23 de Julho de 2025:

Banca Examinadora

Prof. Dr. Eduardo F. de Simas Filho
UFBA (Orientador)

Prof. Dr. Juan Lieber Marin
IFBA (Coorientador)

Prof. Dr. Antônio Carlos Lopes Fernandes Jr.
UFBA

Prof. Dr. Jorge Ricardo de Araujo Kaschny
IFBA

Prof. Dr. Ajalmar Rêgo da Rocha Neto
IFCE

Salvador
2025

Aos meus pais, pelo apoio incondicional e pelos ensinamentos que carrego em cada conquista.

Agradecimentos

Agradeço aos meus pais, Eliana e Jilvan, pelo apoio incondicional e por sempre me incentivarem a seguir em frente.

Aos meus orientadores, Eduardo Simas e Juan Lieber, pela confiança depositada em meu trabalho e por me apresentarem à pesquisa no experimento ATLAS.

Ao professor Eduardo Simas, pela orientação dedicada desde meu trabalho de conclusão de curso. A Juan Lieber, pelo auxílio fundamental nos códigos de implementação no Athena do ATLAS. A Edmar, pela disponibilidade em tirar dúvidas técnicas e compartilhar seu conhecimento sobre o experimento. À colaboração ATLAS, pelo suporte contínuo e pela infraestrutura essencial que possibilitou o desenvolvimento deste trabalho.

Aos amigos que tornaram essa jornada mais leve.

A CAPES pelo suporte financeiro. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

"Remember when you were young?
You shone like the Sun."
Wright, Waters & Gilmour (1975)

Resumo

Na física de altas energias, o grande volume de dados – com uma parcela significativa proveniente de ruído de fundo – dificulta a identificação dos fenômenos de interesse. Para contornar essa complexidade, é adotado um sistema de seleção online de eventos (*trigger*), que no experimento ATLAS do *Large Hadron Collider* (LHC) opera em duas etapas sequenciais: o primeiro nível e o *trigger* de alto nível (também denominado etapa rápida). Técnicas de aprendizado de máquina têm sido empregadas para aprimorar essa seleção.

No contexto da detecção de elétrons, a estimação precisa da energia depositada nos calorímetros é fundamental para a correta identificação de candidatos. Os erros nessa medida podem decorrer de efeitos de *pileup*, que aumentam artificialmente a energia reconstruída, e de perdas longitudinais e laterais de energia, que resultam em uma subestimação da energia verdadeira. Como consequência, imprecisões na estimativa podem comprometer a eficiência da seleção ou aumentar as taxas de falsos positivos.

Este trabalho propõe uma metodologia de calibração de energia para o *trigger* de alto nível, utilizando um *ensemble* de árvores de decisão potenciadas por gradiente (*Gradient Boosted Decision Trees* – GBDTs). Essa abordagem permite modelar não linearidades e capturar relações complexas nos dados de forma eficiente, aprimorando a precisão da estimativa de energia.

A solução desenvolvida foi integrada ao ambiente de software de operação do detector, e está atualmente em avaliação para possível adoção nos próximos anos. Nos testes realizados com dados simulados, observou-se uma redução de até 25% na dispersão da energia reconstruída, além de uma melhora de 20,6% no erro percentual absoluto médio na faixa de baixas energias (0 a 30 GeV). Com dados de validação, foi possível reduzir o limiar de seleção sem comprometer a eficiência na identificação de elétrons, resultando em menor taxa de falsos positivos, menor demanda computacional e aumento da eficiência global do sistema de *trigger*.

Palavras-chave: Física de altas energias, ATLAS, LHC, trigger, calibração de energia, aprendizado de máquina.

Abstract

In high-energy physics, the very large data volume—with a substantial fraction arising from background noise—makes it difficult to identify the phenomena of interest. To overcome this complexity, an online event-selection system (the trigger) is employed, which in the ATLAS experiment at the LHC operates in two sequential stages: the first-level trigger and the High-Level Trigger (also called the rapid stage). Machine-learning techniques have been adopted to enhance this selection.

In the context of electron identification, an accurate estimation of the energy deposited in the calorimeters is crucial for the correct selection of candidates. Measurement errors can stem from pileup effects, which artificially increase the reconstructed energy, and from longitudinal and lateral energy losses, which lead to an underestimation of the true energy. Consequently, inaccuracies in the energy estimate can compromise selection efficiency or raise the false-positive rate.

This work proposes an energy-calibration methodology for the High-Level Trigger using an ensemble of gradient-boosted decision trees (GBDTs). This approach is capable of modeling non-linearities and capturing complex relationships in the data efficiently, thereby improving the precision of the energy estimate.

The solution was integrated into the detector's operational software framework and is currently under evaluation for potential adoption in the coming years. In tests performed with simulated data, we observed up to a 25% reduction in the dispersion of reconstructed energy, as well as a 20.6% improvement in the mean absolute percentage error in the low-energy range (0–30 GeV). With validation data, it was possible to lower the selection threshold without compromising the electron identification efficiency—resulting in a reduced false-positive rate, lower computational demand, and an overall increase in trigger performance.

Keywords: High-energy physics, ATLAS, LHC, trigger, energy calibration, machine learning.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ATLAS *A Toroidal LHC ApparatuS*

CERN *Conseil Européen pour la Recherche Nucléaire*

CMS *Compact Muon Solenoid*

ECAL *Electromagnetic Calorimeter*

EM Eletromagnéticas

GBDT *Gradient Boosting Decision Trees*

HLT *High Level Trigger*

HCAL *Hadronic Calorimeter*

LAr *Liquid Argon*

LGBM *Light Gradient Boosting Machine*

LHC *Large Hadron Collider*

MAE *Mean Absolute Error*

MAPE *Mean Absolute Percentage Error*

MSE *Mean Squared Error*

PMT *Photomultipliers Tubes*

RoI *Region of Interest*

RNA Redes Neurais Artificiais

RSS *Residual Sum of Squares*

TileCal *Tile Calorimeter*

ROC *Receiver Operator Characteristic*

AUC *Area Under the Curve*

Lista de símbolos

η	Pseudo-rapidez.
E	Energia por partícula incidente.
E_T	Energia Transversa.
ϕ	Ângulo azimutal.
ψ	Função base.
$\mathbf{w}$	Vetor de pesos.
λ	Parâmetro de regularização que controla o equilíbrio entre complexidade da árvore e ajuste aos dados.
R_j	Região j da árvore (nó terminal ou folha).
γ_j	Constante de predição (valor do nó terminal) da região R_j em uma árvore de regressão.
Q_m	impureza quadrática da região R_j
Θ	Conjunto de parâmetros da árvore, incluindo todas as regiões e seus valores preditos associados.
ρ	Taxa de aprendizado (ou passo de descida). Escalar positivo que determina o tamanho do passo na direção do gradiente negativo.
r_k	Valor normalizado do anel k.
R_k	Valor do anel K.
$\mathbf{g}$	Gradiente da função de perda L.
$L(\mathbf{f})$	Função de perda (ou função objetivo) que quantifica o erro entre as previsões do modelo e os valores reais.
α	Fator de correção da energia, definido como a razão entre a energia transversa verdadeira do evento e da estimada.
α_{BDT}	Estimador de α dado pelo modelo GBDT.
$< \mu >$	Número médio de interações por colisão.

Sumário

1 Introdução

1.1 Introdução

Os experimentos em física de altas energias utilizam instalações tecnológicas avançadas para medir com precisão as propriedades fundamentais das partículas subatômicas e explorar processos físicos ainda não observados experimentalmente (GRAY, 2021; BETTINI, 2014) . Nesse cenário, aceleradores são frequentemente empregados para provocar colisões entre feixes de partículas. Os subprodutos dessas colisões interagem com sistemas de instrumentação localizados ao redor dos pontos de colisão. Durante essas colisões, uma grande quantidade de processos físicos é gerada, sendo que a maioria deles já é conhecida. Dependendo do processo físico em questão, isso pode ser considerado ruído de fundo.

Para lidar com essa situação, sistemas de filtragem *online*, também conhecidos como *trigger*, são projetados para preservar informações de eventos raros e relevantes, ao mesmo tempo em que eliminam o ruído de fundo. Contudo, em algumas ocasiões, as assinaturas dos diferentes tipos de eventos podem ser bastante similares, o que complica a tarefa de distingui-los.

Recentemente, técnicas de aprendizado de máquina têm sido aplicadas em diversas áreas, como na detecção de partículas, reconstrução de eventos e filtragem *online* (MORAIS et al., 2023; PINTO; ATLAS Collaboration, 2019). Essas técnicas permitem que os sistemas de classificação identifiquem padrões e características nos sinais captados pelos sistemas de instrumentação, aprimorando a distinção entre eventos de interesse e o ruído de fundo.

O LHC, localizado na fronteira entre a França e a Suíça, é um dos aceleradores mais importantes na física de partículas, atualmente. Ele acelera feixes de prótons, em direções opostas, em um túnel circular de aproximadamente 27 km de extensão, com colisões ocorrendo a cada 25 nanosegundos (BRÜNING; BURKHARDT; MYERS, 2012). O experimento *A Toroidal LHC ApparatuS* (ATLAS), situado em um dos pontos de colisão do LHC, é composto por vários subdetectores e gera cerca de 1,3 MB de dados por colisão (EVANS; BRYANT, 2008). Com a taxa de colisão nominal do LHC, o ATLAS produz aproximadamente 52 TB de dados por segundo, o que requer um sistema de seleção online (*trigger*) para identificar os eventos relevantes que são armazenados para análises posteriores (ORELLANA, 2020).

Dentro dos subdetectores do ATLAS, os calorímetros são sistemas de instrumentação eletrônica dedicados à medição da energia das partículas, sendo especialmente importantes para o sistema de *trigger online* devido à sua resposta rápida e às características distintivas nos perfis de deposição de energia, essenciais para a detecção de eventos de interesse (WIGMANS, 2017). No ATLAS, esses calorímetros são finamente segmentados em sete camadas sobrepostas, com aproximadamente 182.500 canais de leitura (SOTO, 2025).

Os procedimentos de calibração de energia realizados tanto nas etapas do *High Level Trigger* (HLT) quanto na análise *offline* dos eventos aprovados pelo *trigger* são essenciais para caracterizar corretamente as assinaturas de interesse. Atualmente, essas etapas de calibração, que

aumentam consideravelmente a resolução da medição dos calorímetros, estão disponíveis apenas no processamento *offline* (WIGMANS, 2017; ATLAS Collaboration, 2017b). Em particular, para partículas de natureza eletromagnética, como elétrons, pósitrons e fótons, a energia é estimada através da soma das informações das células das camadas eletromagnéticas do calorímetro, que estão contidas em uma região de interesse (*Region of Interest* (RoI)) ao redor da célula com a maior deposição de energia. Esse procedimento, embora eficiente em termos computacionais, pode introduzir erros de medição devido a perdas laterais ou longitudinais fora da RoI, ou ainda à contaminação por energia oriunda de interações simultâneas (efeito de *pileup*). Tais imprecisões podem comprometer o desempenho do sistema de seleção online, impactando diretamente a eficiência e a taxa de falsos positivos no *trigger* (MARIN, 2020).

1.2 Justificativa

No experimento ATLAS, a etapa rápida (*fast step*) do sistema de *trigger* de alto nível (HLT) desempenha um papel crucial na seleção de eventos relevantes para análise posterior. No entanto, a calibração de energia nesta etapa enfrenta desafios significativos devido à resolução limitada dos calorímetros e ao curto intervalo de tempo disponível nessa etapa. Atualmente, os métodos de calibração mais precisos estão disponíveis apenas nas etapas *offline* e no final da sequência de algoritmos do sistema *online* do *trigger* (ATLAS Collaboration, 2024), deixando uma lacuna na etapa rápida onde a seleção inicial do HLT é realizada.

A proposta desta dissertação foi implementar um método de calibração de energia na etapa rápida do *trigger*. A calibração proposta visa melhorar a precisão na estimativa de energia dos elétrons, o que pode reduzir a taxa de falsos positivos (eventos irrelevantes erroneamente classificados como significativos). A redução desses falsos positivos é crucial para evitar a sobrecarga no processamento e a análise excessiva de dados não relevantes (ATLAS Collaboration, 2022).

Ao aplicar um método de calibração mais avançado, espera-se melhorar a precisão da filtragem na etapa rápida, garantindo que o sistema de *trigger* selecione eventos relevantes com maior eficácia. Isso não só reduz a quantidade de dados não significativos processados nas fases subsequentes, mas também aumenta a eficiência geral do sistema. O que seria benéfico para a operação eficiente de um experimento de grande escala, como o ATLAS, onde o processamento de grandes volumes de dados é um desafio contínuo. Otimizando os requisitos computacionais e contribuindo para o sucesso do experimento. A proposta de um novo método de calibração visa aprimorar a qualidade dos dados armazenados e aumentar a confiabilidade das análises subsequentes.

1.3 Objetivo

O principal objetivo desta pesquisa foi projetar, avaliar e implementar um algoritmo de aprendizado de máquina para a calibração de energia dos calorímetros na etapa rápida do sistema de seleção online do experimento ATLAS, visando aprimorar a identificação de elétrons. Para alcançar esse objetivo, foram definidos os seguintes objetivos específicos:

- Treinamento do algoritmo de calibração utilizando um conjunto de árvores de decisão com reforço por gradiente, com dados provenientes das simulações de Monte Carlo da Run 3 do LHC (2022 - 2026);
- Análise dos resultados obtidos durante o treinamento;
- Implementação do modelo de calibração no software operacional do ATLAS (Athena);
- Avaliação da eficiência do algoritmo nas cadeias de seleção *online* do ATLAS com dados simulados e experimentais.

1.4 Organização do Documento

No capítulo 2, são apresentados os fundamentos teóricos relacionados à física de partículas e ao experimento ATLAS, com foco nos sistemas de calorimetria e no sistema de seleção online de elétrons. No capítulo 3, fundamenta-se o aprendizado de máquina e as árvores de decisão, com ênfase na árvore de decisão reforçada por gradiente.

No Capítulo 4, é apresentada a metodologia utilizada durante o desenvolvimento do trabalho, os dados utilizados e as medidas de avaliação de desempenho para os métodos propostos. No Capítulo 5, discutem-se os resultados obtidos com o algoritmo proposto para a calibração de energia e a comparação com o método atual utilizado pelo ATLAS. Finalmente, o Capítulo 6 apresenta a conclusão do trabalho, com uma síntese crítica dos resultados obtidos, reflexões sobre as limitações identificadas, e perspectivas para aplicações ou aprimoramentos futuros na calibração de energia em experimentos de física de partículas. As publicações e materiais relacionados a este trabalho estão reunidos no Apêndice A.

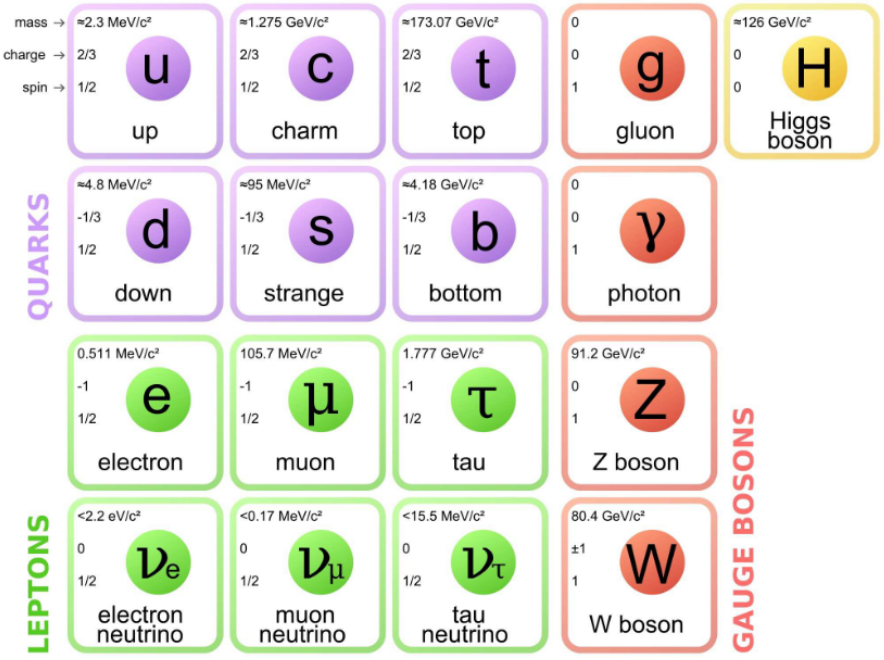
2 Física de Partículas e o experimento ATLAS

2.1 Física de Partículas

A Física de Partículas é um ramo da física que busca entender a natureza fundamental da matéria e das forças que governam o universo em escalas subatômicas. Desde os primórdios da humanidade, o homem tem buscado compreender a estrutura fundamental do cosmos e as partículas que o compõem. No entanto, foi apenas no século XX que a física de partículas emergiu como um campo científico distinto, impulsionado por avanços tecnológicos e teóricos (ARBUZOV, 2018).

Atualmente, no cerne da física de partículas está o Modelo Padrão (ARBUZOV, 2018), Figura 1, um arcabouço teórico que descreve as partículas fundamentais e as interações fundamentais que governam o universo. O Modelo Padrão é uma conquista notável da física moderna, unificando três das quatro forças fundamentais da natureza: eletromagnetismo, interação fraca e interação forte, com exceção da gravidade. As partículas fundamentais descritas pelo Modelo Padrão incluem férmions (quarks e léptons), bósons de gauge (fótons, bósons W e Z, bósons gluônicos) e o recentemente descoberto, em 2012, bóson de Higgs (ATLAS Collaboration, 2012).

Figura 1 – Modelo padrão da física de partículas.



Fonte: (ARBUZOV, 2018)

A compreensão e validação experimental do Modelo Padrão foram alcançadas através de uma série de experimentos de alta energia realizados em aceleradores de partículas em todo o mundo, como o LHC ou grande colisor de hádrons no *Conseil Européen pour la Recherche Nucléaire* (CERN) ou em português (Organização Europeia para a Pesquisa Nuclear). Esses

aceleradores permitem que os físicos colidam partículas com energias extremamente altas, fornecendo *insights* sobre as propriedades das partículas elementares e as leis fundamentais que regem sua interação.

Além de elucidar a estrutura fundamental da matéria, a física de partículas também desempenha um papel crucial na compreensão da evolução do universo. Teorias como a inflação cósmica e a matéria escura são investigadas através de experimentos de física de partículas, oferecendo novas percepções sobre a história e a composição do cosmos.

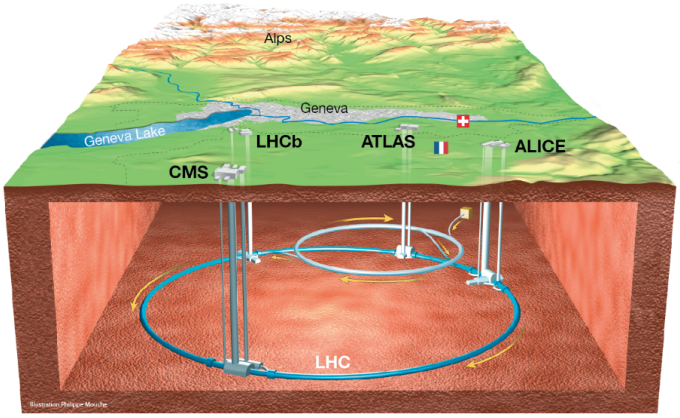
O LHC foi construído entre 1998 e 2008 nas instalações do CERN (EVANS; BRYANT, 2008). Localizado na fronteira entre a França e a Suíça, é um acelerador de partículas projetado para colidir feixes de prótons ou íons pesados, proporcionando a oportunidade de observar eventos raros e detectar partículas de interesse para a física de partículas.

O LHC foi concebido para gerar colisões de partículas com energia no centro de massa de até 14 TeV, a uma taxa de 40 $\times$ 10 eventos por segundo. Os feixes percorrem o anel em sentidos opostos e se cruzam em pontos definidos, onde detectores foram instalados para analisar as propriedades das partículas produzidas. São quatro os detectores principais: ATLAS e CMS (experimentos de propósito geral), ALICE (focado em colisões entre núcleos) e LHCb (especializado em física de quarks b) (EVANS; BRYANT, 2008). Conforme ilustrado na Figura 2, esse acelerador ocupa um túnel circular situado a cerca de 100 metros abaixo da superfície, com um percurso total de 27 km.

Ele desempenha um papel crucial na validação experimental do Modelo Padrão da física de partículas e na busca por novas partículas e fenômenos físicos além deste modelo teórico. Sua construção e operação representam uma contribuição significativa para o avanço do conhecimento científico no campo da física de partículas.

Atualmente, na Run 3 (2022-2026), as colisões estão sendo realizadas em energias recordes, proporcionando uma oportunidade sem precedentes para explorar novos fenômenos e validar teorias físicas.

Figura 2 – Instalações do LHC e seus experimentos.

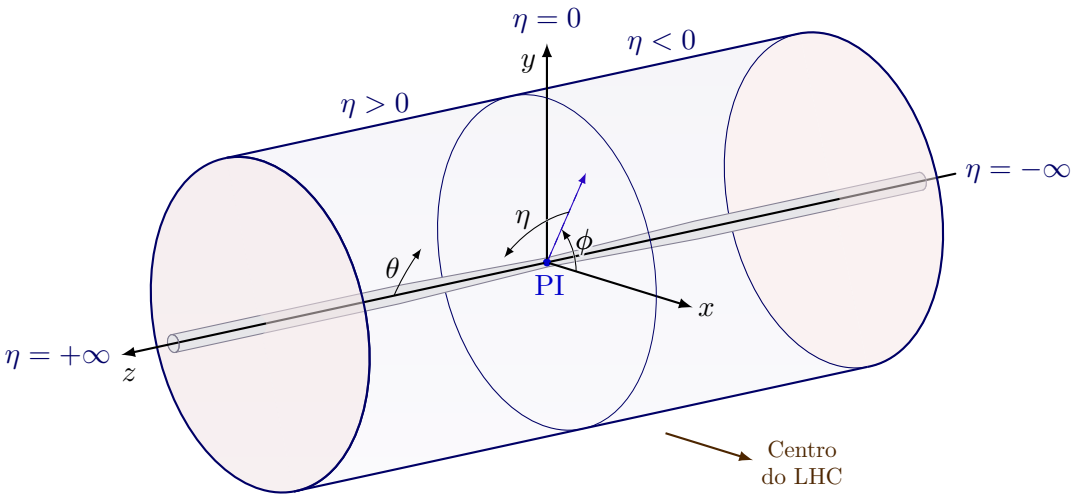


Fonte: (EVANS; BRYANT, 2008)

2.2 Experimento ATLAS

O ATLAS é um dos experimentos que fazem parte do LHC, e em 2012 foi responsável, juntamente com o *Compact Muon Solenoid* (CMS), por comprovar a existência do bóson de Higgs (ATLAS Collaboration, 2012). No ATLAS é utilizado um sistema de coordenadas cilíndricas para descrever as trajetórias das partículas produzidas nas colisões, conforme é possível observar na Figura 3. O eixo x é direcionado a partir do ponto de interação (PI) até o centro do anel do LHC e o eixo y é perpendicular a este plano. As coordenadas cilíndricas (r, ϕ) são usadas no plano transverso, com ϕ sendo o ângulo azimutal em torno do eixo z.

Figura 3 – Esboço do sistema de coordenadas do ATLAS



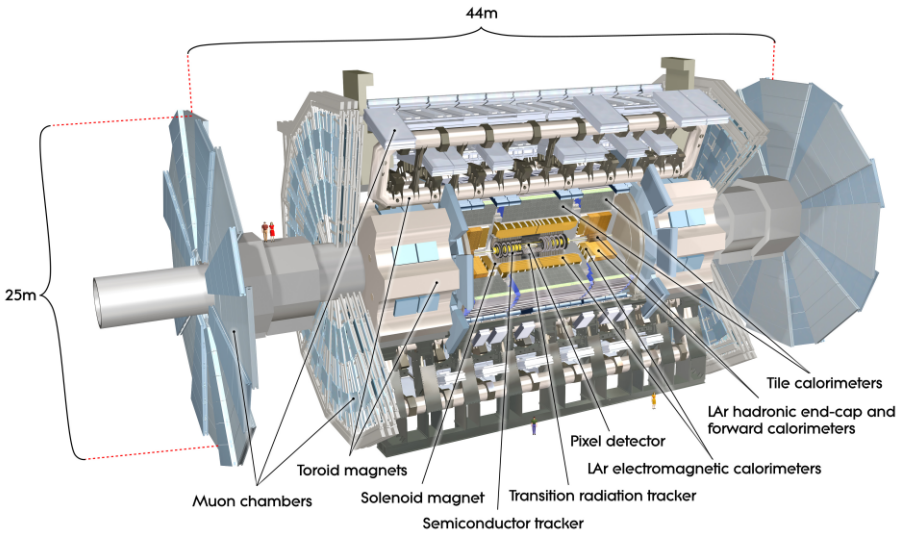
Fonte: Autoria Própria

Define-se a pseudo rapidez η em termos do ânguo polar θ como:

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

A pseudo-rapidez, η, determina a elevação da partícula em relação ao eixo z. Por exemplo, $\eta = 0$ determina uma linha perpendicular ao eixo z enquanto que quão maior for o valor de η mais próximo da linha do feixe estará a coordenada, de tal forma, o limite $\eta \to \infty$ representa uma coordenada completamente paralela ao feixe. Como apresentado na Figura 4 o experimento possui aproximadamente 25 m de altura e 44 m de comprimento e consiste em um detector interno rodeado por um solenoide supercondutor fino, calorímetros eletromagnético e hadrônico e um espectrômetro de múons (ATLAS Collaboration, 2008). Na física de partículas, devido à geometria cilíndrica dos detectores nos colisores de partículas, a energia depositada é medida sobretudo no plano perpendicular ao eixo do feixe. Por isso, nas análises utiliza-se a energia transversal, definida como: $E_T = E \sin\theta$.

Figura 4 – Detector ATLAS e seus subsistemas



Fonte: (ATLAS Collaboration, 2017c)

Além disso, devido a quantidade de informação gerada, em torno de 52 TB/s, faz-se necessário a utilização de um sistema de *online* de seleção de eventos (*trigger*), de acordo com o evento de interesse para os pesquisadores, visando selecionar quais dados serão salvos para análise *offline*. O sistema é composto por dois níveis o L1 (do inglês *Level* 1) e o HLT (do inglês *High Level Trigger*). O primeiro nível é implementado em *hardware* customizado e tem como objetivo reduzir a taxa de eventos de 40 MHz para 100 kHz em 2.5 μs. Também define as regiões de interesse (do inglês *RoI*) que tem *clusters* do calorímetro com alta energia transversa E_T. Os eventos aceitos por L1 são processados no segundo nível, HLT, baseados em algoritmos implementados em *software* que reduzem ainda mais a taxa de eventos salvos em disco, para 1 kHz (ATLAS Collaboration, 2020).

2.2.1 Calorimetria no ATLAS

Em física de partículas, calorimetria se refere à detecção de partículas por meio da interação dessas com a matéria, de forma que a medição ocorre a partir da absorção total, dessa forma, o processo é em geral destrutivo e as partículas não estarão disponíveis após sua passagem pelo calorímetro.

É possível classificar os calorímetros quanto a sua composição em homogêneos ou amostradores (WIGMANS, 2017). Nos calorímetros homogêneos, todo o volume de seu material é sensível a partícula incidente e consequentemente contribui para a produção do sinal. Já os calorímetros amostradores possuem um material passivo responsável por absorver a energia da partícula incidente, e um material ativo para produção do sinal utilizado na detecção.

As interações com o calorímetro ocorrem de forma diferente para os diferentes tipos de partículas. Partículas Eletromagnéticas (EM), como elétrons e posítrons, necessitam de pequena quantidade de material para serem absorvidas, pois geram um chuveiro de partículas menos

energéticas ao interagir com a matéria. Já os múons, necessitam de grande quantidade de material para sua absorção, pois perdem energia lentamente. As partículas hadrônicas, que interagem por meio da força nuclear forte, geram chuveiros mais complexos de descrever em comparação com os produzidos por partículas eletromagnéticas, devido à maior diversidade de processos envolvidos (WIGMANS, 2017).

Devido a diferença de características são utilizados calorímetros específicos para estas partículas. O calorímetro eletromagnético é usualmente instalado internamente ao hadrônico. Visto que, as partículas eletromagnéticas são, tipicamente, absorvidas nas camadas mais internas. Os chuveiros hadrônicos, por sua vez, só são completamente absorvidos nas camadas hadrônicas, mais externas.

A precisão dos calorímetros, diferente de outros tipos de detectores, aumenta com a energia (WIGMANS, 2017), de acordo com a Eq. (2.1).

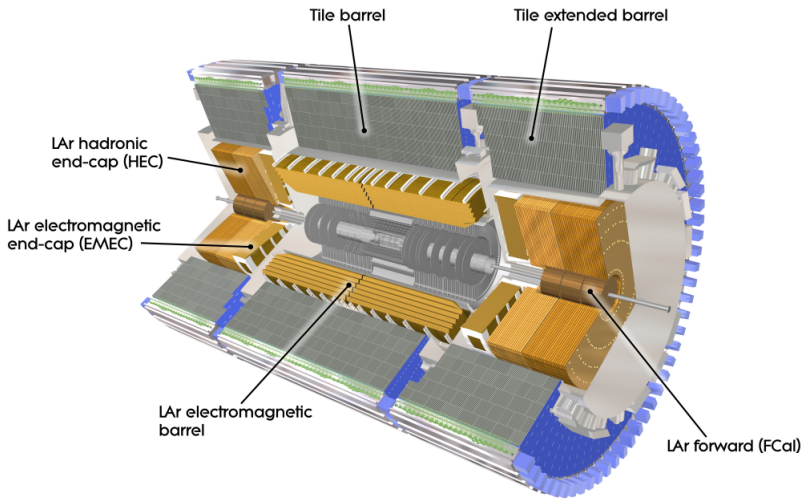
$$\frac{\sigma_E}{E} \propto \frac{1}{\sqrt{E}}, \tag{2.1}$$

em que E é a energia por partícula incidente e σ_E é o desvio padrão esperado na medição.

O calorímetro do detector ATLAS é composto por 7 camadas (EVANS; BRYANT, 2008). Sendo 4 eletromagnéticas (PS, E1, E2, E3) e 3 hadrônicas (H0, H1 e H2). A sua parte eletromagnética é também chamada de *Electromagnetic Calorimeter* (ECAL), enquanto a hadrônica de *Hadronic Calorimeter* (HCAL).

Cada uma dessas camadas apresenta diferentes concentrações de células detectoras por unidade de área (granularidade). O calorímetro é disposto de forma simétrica e com cobertura total em torno do ângulo azimutal (ϕ) e cobrindo a faixa de pseudo-rapidez $\mid \eta \mid < 4{,}9$. É possível ver uma ilustração do sistema na Figura 5.

Figura 5 – Ilustração do sistema de calorimetria do ATLAS.

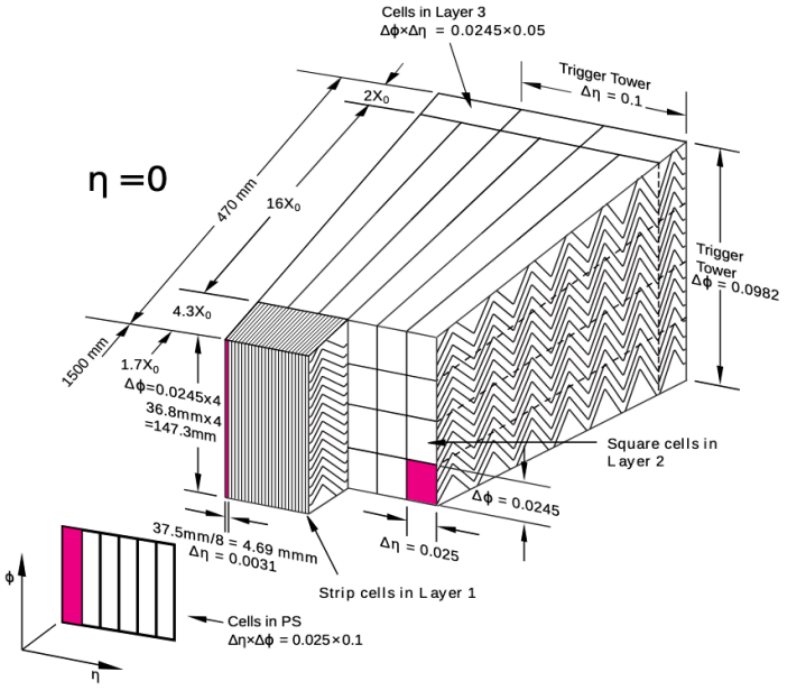


Fonte: (EVANS; BRYANT, 2008)

A região central do detector, chamada de barril (do inglês *barrel*), é dividida em duas metades (-1,475 $< \eta <$ 0 e 0 $< \eta <$ 1,475) e em suas extremidades ficam as a tampas (do inglês, *end-caps* que cobrem a região de 1,375 $<$ | η | $<$ 3,2. Na região de transição entre o barril e as tampas (1,37 $<$ | η | $<$ 1,54) a quantidade de material ativo é reduzida devido à passagem de cabos, conexões e fibras óticas, dificultando a caracterização das partículas que atravessam o detector naquela vizinhança. Por essa razão, muitas análises físicas optam por remover dados dessa região. Já na região entre 2,47 $<$ | η | $<$ 2,5 existe uma junção dos dois blocos cilíndricos que compõem as tampas mais interna e externa do detector (conhecidas como barril estendido, do inglês *extended barrel*). Essa região tem uma menor quantidade de sensores devido a presença de suportes mecânicos.

O calorímetro eletromagnético é composto por argônio líquido, *Liquid Argon* (LAr), as três camadas que o compõe (EM1, EM2 e EM3) compartilham do mesmo eixo da estrutura cilíndrica, e foram concebidos com o objetivo de recuperar as informações de desenvolvimento longitudinal da parcela eletromagnética dos chuveiros das partículas que o atravessam. Na Figura 6, pode-se ver uma ilustração da disposição de células do ECAL nas três camadas. Na primeira camada, as células possuem geometria retangular, com maior dimensão paralela ao eixo η. Já na EM2 as células são aproximadamente quadradas, nessa camada há a maior captura da fração do chuveiro eletromagnético, visto que possui a maior profundidade. A última camada é responsável pela captura da cauda do chuveiro, apresentando células que possuem maior dimensão no plano $\eta \times \phi$ (EVANS; BRYANT, 2008). Antes do calorímetro EM temos ainda um sistema pré-amostrador (do inglês *PS - Pre Sampler*), que cobre o intervalo de | η |$<$ 1.8 e é utilizado para estimar a perda de energia das partículas antes de chegarem nos calorímetros.

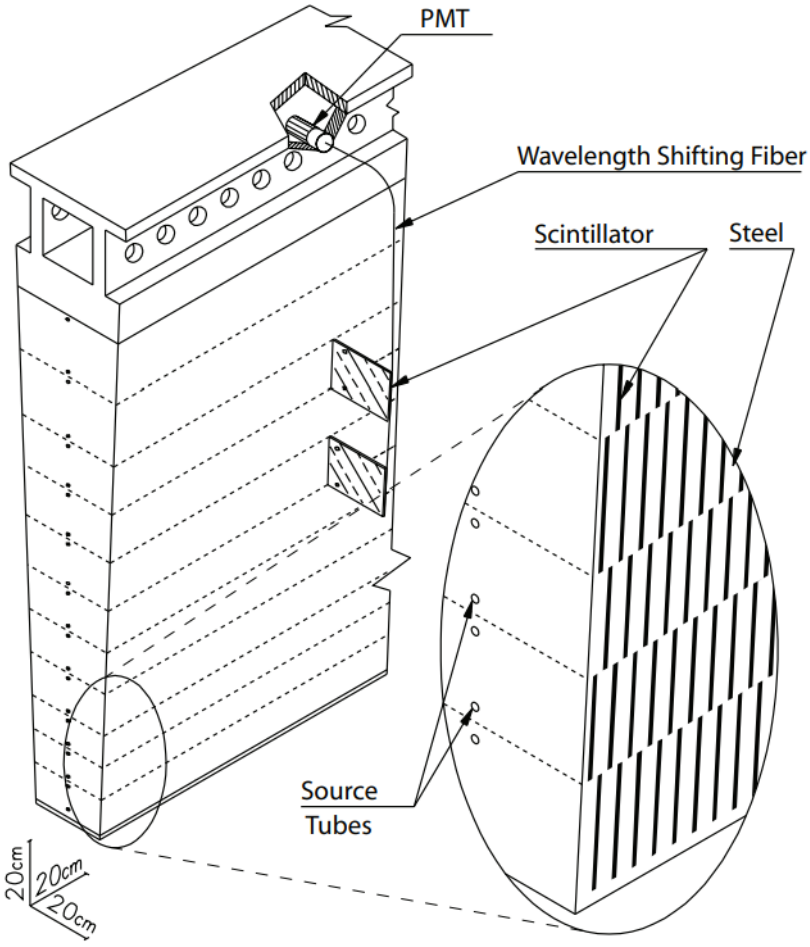
Figura 6 – Representação da granularidade das camadas do barril do ECAL em $\eta = 0$.



Fonte: (ATLAS Collaboration, 2017a)

Já o calorímetro hadrônico de telhas, *Tile Calorimeter* (TileCal), circunda o calorímetro EM e é utilizado para medir as partículas de natureza hadrônica, como prótons e píons (EVANS; BRYANT, 2008).O TileCal é um calorímetro de amostragem que utiliza aço como material absorvedor e telhas cintiladoras como material ativo, os sinais são gerados a partir de dispositivos fotomultiplicadores, *Photomultipliers Tubes* (PMT), que convertem luz em sinal elétrico. O diagrama de um único módulo do TileCal pode ser visto na Figura 7. A amplitude do sinal gerado nos canais do TileCal é proporcional a energia da partícula que atravessa seu material.

Figura 7 – Diagrama esquemático de um módulo do TileCal



Fonte: (ATLAS Collaboration, 1997)

Para o ECAL, as células responsáveis por medir a energia das partículas que atravessam o calorímetro do ATLAS possuem tamanhos diferentes dependendo da coordenada η em que estão localizadas. As variações de granularidade de cada camada do sistema no plano $\eta \times \phi$ podem ser observadas na Tabela 1.

Tabela 1 – Granularidade das células e camadas utilizadas no ATLAS.

Camada	Amostrador		Cobertura	Granularidade ($\Delta\eta \times \Delta\phi$)
Pré-amostrador	Barril	PSB	$0,00 < \|\eta\| < 1.58$	0,025×0,1
	Tampa	PSE	$1,50 < \|\eta\| < 1,80$	0,025×0,1
Calorímetro Eletromagnético				
Camada 1	Barril	EMB1	$0,00 < \|\eta\| < 1.55$	0,003×0,1
	Tampa	EMEC1	$1,37 < \|\eta\| < 1.80$	0,003×0,1
			$1,80 < \|\eta\| < 2,00$	0,025×0,1
			$2,00 < \|\eta\| < 2,37$	0,006×0,1
			$2,37 < \|\eta\| < 3,20$	0,1×0,1
Camada 2	Barril	EMB2	$0,00 < \|\eta\| < 1,50$	0,025×0,025
	Tampa	EMEC2	$1,35 < \|\eta\| < 2,50$	0,025×0,025
			$2,50 < \|\eta\| < 3,20$	0,1×0,1
Camada 3	Barril	EMB3	$0,00 < \|\eta\| < 1.58$	0,05×0,1
	Tampa	EMEC3	$1,35 < \|\eta\| < 2,50$	0,05×0,025
			$2,50 < \|\eta\| < 3,20$	0,1×0,1
Calorímetro Hadrônico				
Camada 1	Barril	TileCal1	$0,00 < \|\eta\| < 1,09$	0,1×0,1
	Barril Extendido	TileExt1	$0,94 < \|\eta\| < 1,77$	0,1×0,1
	Tampa	HEC1	$1,50 < \|\eta\| < 2,50$	0,1×0,1
			$2,50 < \|\eta\| < 3,20$	0,2×0,2
Camada 2	Barril	TileCal2	$0,00 < \|\eta\| < 1,09$	0,1×0,1
	Barril Extendido	TileExt2	$0,85 < \|\eta\| < 1,41$	0,1×0,1
	Tampa	HEC2	$1,50 < \|\eta\| < 2,50$	0,1×0,1
			$2,50 < \|\eta\| < 3,20$	0,2×0,2
Camada 3	Barril	TileCal3	$0,85 < \|\eta\| < 0,72$	0,2×0,1
	Barril Extendido	TileExt3	$0,85 < \|\eta\| < 1,41$	0,2×0,1
	Tampa	HEC3	$1,50 < \|\eta\| < 2,50$	0,1×0,1
			$2,50 < \|\eta\| < 3,20$	0,2×0,2

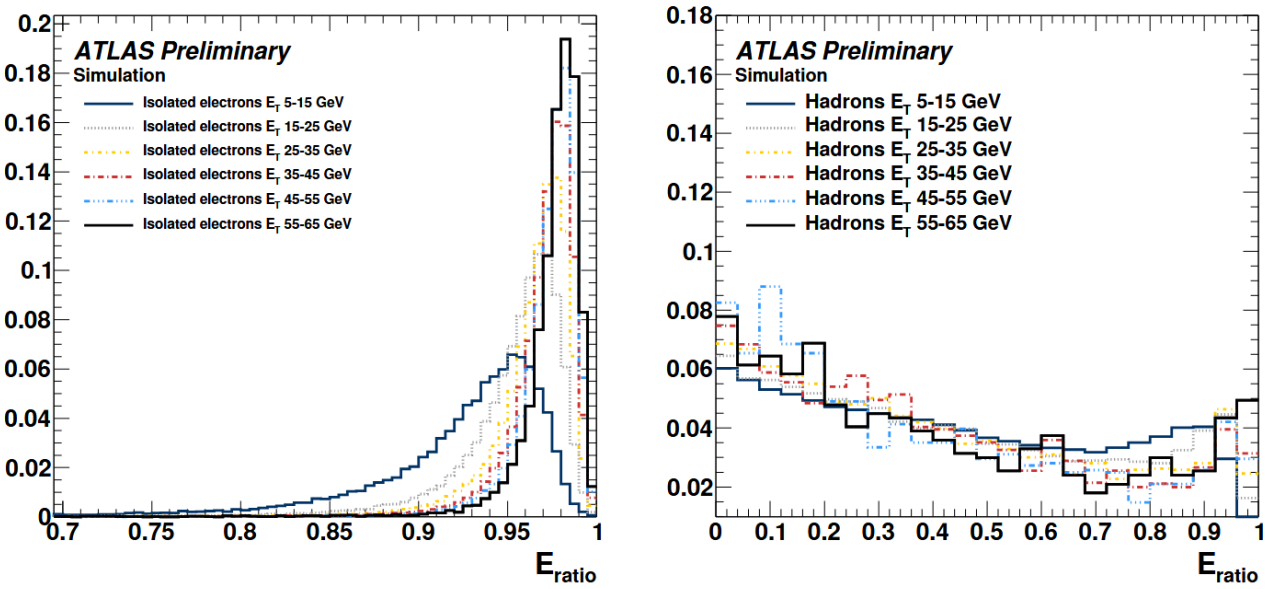
Fonte: (EVANS; BRYANT, 2008)

Devido à não simetria e à granularidade variável do calorímetro na coordenada η, os perfis de deposição de energia diferem de região para região. Além disso, esses perfis mudam conforme a energia transversal E_T da partícula. É possível observar essa dependência na Figura 8, onde a variável

$$E_{\text{ratio}} = \frac{E_{\text{max}} - E_{\text{2nd}}}{E_{\text{max}} + E_{\text{2nd}}},$$

com E_{max} e E_{2nd} o primeiro e o segundo pico de energia transversal na primeira camada do ECAL. Os histogramas de E_{ratio} mostram claramente como o formato do chuveiro se torna mais ou menos concentrado à medida que E_T varia. Por conta dessas variações regionais (em η) e energéticas (em E_T), justifica-se uma segmentação do espaço de fase em *bins* de η e E_T, de modo que cada modelo treinado possa aprender e reproduzir com maior precisão as características da resposta calorimétrica em cada região e faixa energética.

Figura 8 – Razão da diferença de energia associada ao maior e ao segundo maior depósito de energia sobre a soma dessas energias para elétrons isolados (à esquerda) e hádrons (à direita) em vários intervalos de E_T.



Fonte: (ATLAS Collaboration, 2011)

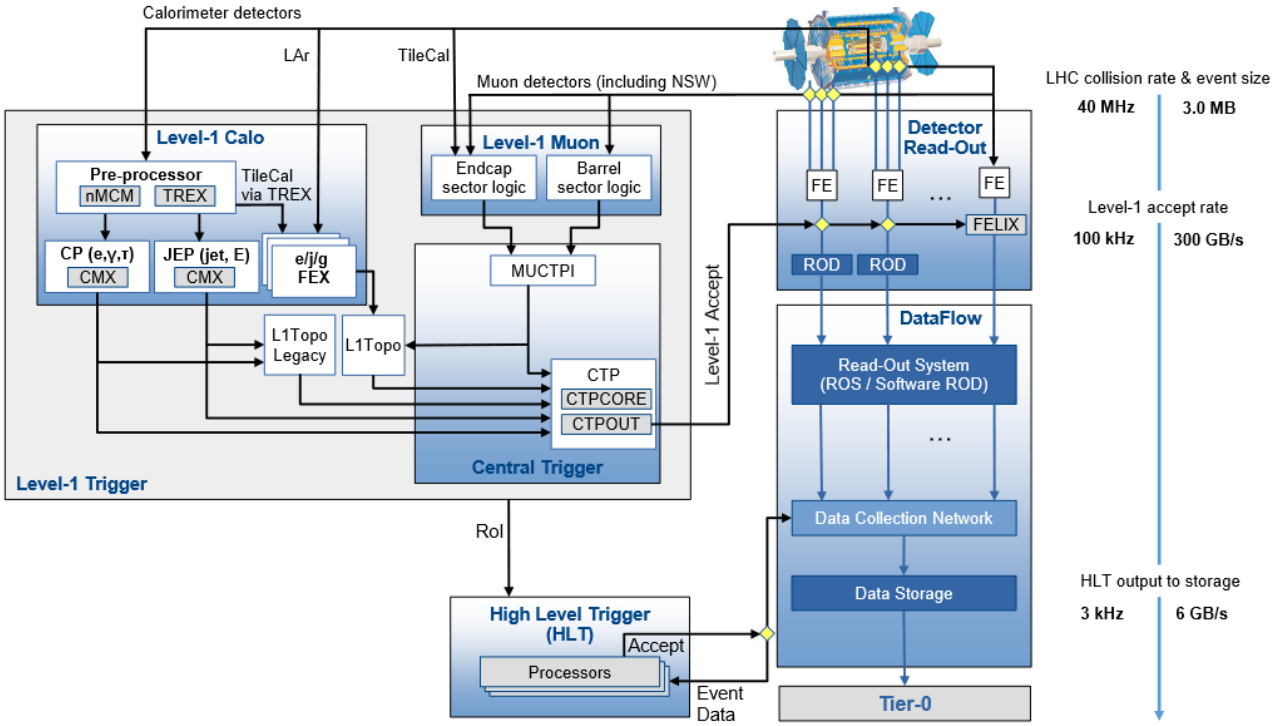
2.2.2 Filtragem *Online* de Elétrons

Como mencionado anteriormente o experimento ATLAS utiliza um sistema de *trigger* de dois níveis, sendo o primeiro, L1 *trigger*, basedo em *hardware*, que utiliza sistemas eletrônicos dedicados para selecionar as informações provenientes do calorímetro e dos detectores de múons. Esse primeiro nível reduz a taxa de eventos de 40 MHz para 100 kHz, dentro de um espaço de tempo de 2,5 µs. Além disso, também seleciona as regiões do calorímetro que podem conter informações relevantes (RoI - *Region of Interest*) para a identificação da partícula e as transmite para o *trigger* de alto nível (HLT) (ATLAS Collaboration, 2024).

O ATLAS passou, em 2022, por uma atualização substancial com melhorias em vários subsistemas do detector e sua eletrônica, a fim de viabilizar o amplo programa de física planejado para a coleta de dados do *Run 3*.

A Figura 9 apresenta uma ilustração dos componentes internos do L1 bem como o trafégo de dados entre o L1 e o HLT, e a redução da quantidade de informação a cada etapa.

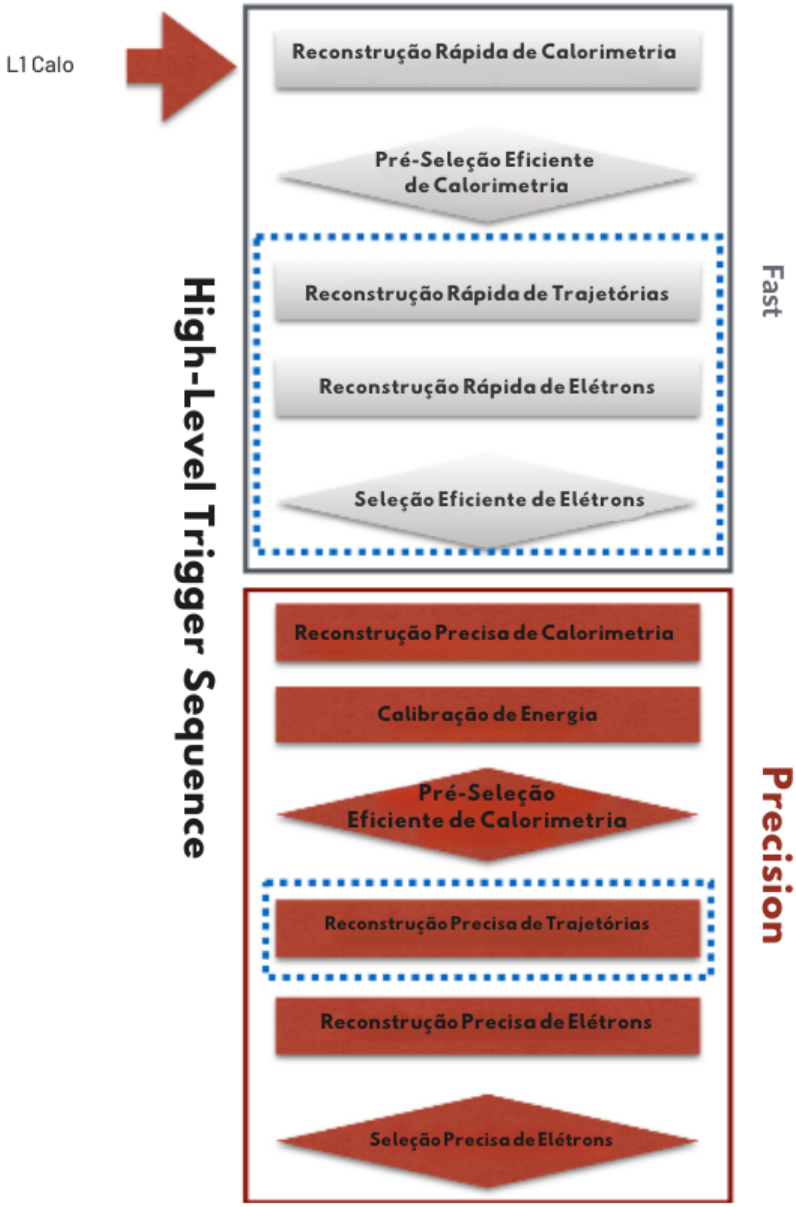
Figura 9 – Sistema ATLAS TDAQ na Run3 com ênfase nos componentes relevantes para o *trigger*, bem como a leitura do detector e o fluxo de dados



Fonte: (ATLAS Collaboration, 2024)

Em posse das RoIs, o segundo nível de filtragem, a etapa HLT, baseada em *software*, realiza a detecção de elétrons em cada RoI EM fornecida pelo L1 que satisfaça o requisito de E_T ou qualquer outro especificado pelo menu de *trigger* (ATLAS Collaboration, 2024). As etapas para a identificação de elétrons podem ser vistas na Figura 10, para elétrons o HLT é dividido em etapa rápida e precisa, sendo a segunda mais rigorosa que a primeira devido à restrições de tempo. A primeira etapa aplica cortes lineares sobre variáveis calorimétricas com elevado poder de discriminação, como a razão de energia hadrônica e o espalhamento lateral do chuveiro eletromagnético, entre outras (ATLAS Collaboration, 2019). A decisão tomada nesta fase rápida é então encaminhada para a etapa de precisão, na qual são empregados algoritmos de seleção de eletróns ainda mais restritivos, além da aplicação da calibração da energia da partícula.

Figura 10 – Sequência de algoritmos do *trigger* de elétrons



Adaptado de: (JONES, 2019)

A reconstrução dos candidatos a elétrons utiliza um algoritmo de janela deslizante com janelas de agrupamento retangulares de tamanho $\Delta\eta \times \Delta\phi = 0{,}075 \times 0{,}175$ no barril e $0{,}125 \times 0{,}125$ nas tampas das extremidades (ATLAS Collaboration, 2024).

A etapa rápida de seleção do calorímetro possui três implementações. A implementação padrão usa um algoritmo baseado em redes neurais, o algoritmo Ringer (PINTO; ATLAS Collaboration, 2019), que utiliza como entrada somas de energia de todas as células em 100 anéis concêntricos centrados ao redor da célula mais energética em cada camada de amostragem do calorímetro. Na Run 2, esse algoritmo foi utilizado apenas para selecionar elétrons com $E_T \geq 15$ GeV [1], mas na

[1] A unidade eV (elétron-volt) é normalmente utilizada nos estudos de física de partículas, em que 1 eV compreende a energia necessária para aumentar o potencial elétrico de um elétron em um volt, equivalente a $1{,}6 \times 10^{-19}$ J

Run 3 está sendo aplicado a partir de $E_T \geq 5$ GeV. O algoritmo Ringer é otimizado em duas regiões de E_T: entre 5–15 GeV com amostras Monte Carlo (MC) de $J/\psi \to e^+e^-$ e para faixas de energia $E_T > 15$ GeV com amostras MC de $Z \to e^+e^-$ (ATLAS Collaboration, 2024).

Para elétrons, é possível também utilizar "seleções rápidas"do calorímetro que utilizam como entrada ou apenas a E_T (baseado em E_T) do *cluster* ou a E_T do *cluster* com três parâmetros de forma de chuveiro (baseado em cortes) (ATLAS Collaboration, 2022).

O limiar de decisão definido por E_T determina a nomenclatura das cadeias do *trigger*: por exemplo, um limiar de $E_T > 22$ GeV origina a cadeia denominada EM22. Esses limiares possuem uma tolerância operacional de -2 a 3 GeV em relação ao valor nominal, projetada para compensar variações na calibração da energia medida nas células do detector.

O HLT opera por meio do *framework* Athena, software central do ATLAS para processamento e análise de dados. Desenvolvido sobre o Gaudi (BARRAND et al., 2001) – *framework* especializado em experimentos de física de altas energias –, o Athena emprega uma arquitetura orientada a componentes, onde algoritmos, serviços e ferramentas são configuráveis dinamicamente via propriedades reajustáveis durante a execução. Essa modularidade garante flexibilidade para adaptação a diferentes cenários de filtragem, essencial para a variedade de cadeias de *trigger*.

Neste trabalho, será feita uma análise sobre as cadeias E26, E60, que correspondem a limiares de E_T de 26 GeV, 60 GeV, respectivamente. A escolha dessas cadeias permite avaliar a estabilidade do sistema de *trigger* em diferentes regimes de energia.

2.2.2.1 Algoritmo *NeuralRinger*

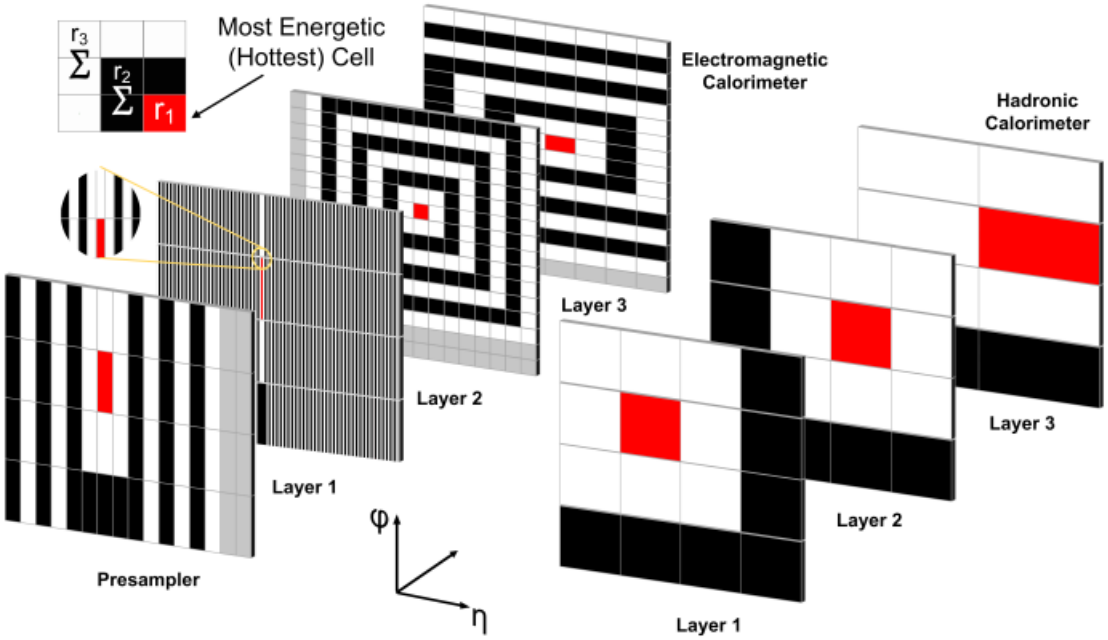
O NeuralRinger é o algoritmo que desde 2017 opera nas cadeias de identificação *online* de elétrons no ATLAS (PINTO; ATLAS Collaboration, 2019), o algoritmo é um *ensemble* de Redes Neurais Artificiais (RNA) que fazem uso da informação proveniente da soma de energia de todas as células contidas em um anel concêntrico (ou *rings*). Os anéis são estruturas que exploram a propriedade dos chuveiros eletromagnéticos de se desenvolverem em uma estrutura aproximadamente cônica ao longo do eixo da interação inicial. Esta característica possibilita a codificação das informações relevantes do chuveiro em quantidades que descrevem a soma de energia de todas as células contidas em um anel concêntrico, gerados em cada camada do calorímetro (ATLAS Collaboration, 2019).

Os anéis têm formato retangular devido a estrutura de células do calorímetro como pode ser visto na Figura 11, o algoritmo de montagem destes opera, no ECAL, mapeando a energia de cada célula do calorímetro em anéis concêntricos, usando como referência a célula mais energética de cada camada em uma janela de $\eta \times \phi = 0{,}4 \times 0{,}4$ (SEIXAS et al., 1996).

No calorímetro EM, os anéis são centrados em torno da célula mais energética de cada camada, enquanto no calorímetro hadrônico, a célula central é definida na mesma posição $\eta \times \phi$ da célula mais energética da segunda camada EM.

no Sistema Internacional de Unidades e Medidas (SI).

Figura 11 – Ilustração da montagem dos anéis do algoritmo NeuralRinger.



Fonte: (ATLAS Collaboration, 2020)

Na Tabela 2 é mostrado o número de anéis em cada camada do detector. Os números são diferentes para cada camada devido a diferença de granularidade das células.

Tabela 2 – Quantidade de anéis por camada do sistema de calorimetria do experimento ATLAS

Camadas	PS	EM1	EM2	EM3	H1	H2	H3
Anéis	8	64	8	8	4	4	4

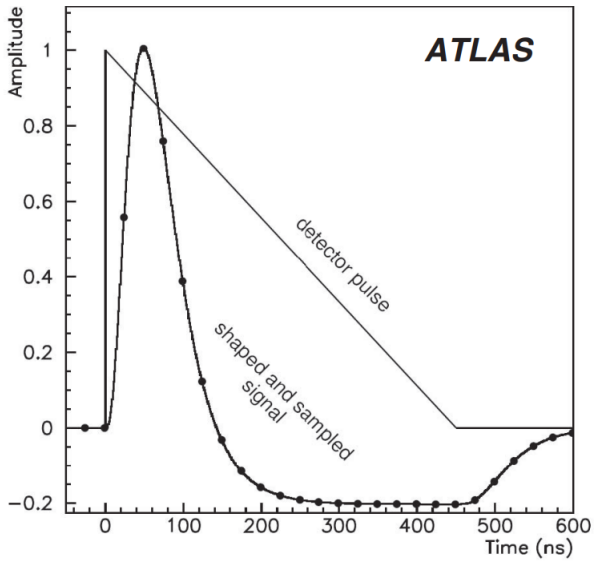
Fonte: (ATLAS Collaboration, 2020)

2.2.3 Estimação e Calibração de Energia

Como descrito anteriormente, partículas carregadas em chuveiros eletromagnéticos ou hadrônicos que se desenvolvem no calorímetro EM ionizam o argônio líquido ao atravessarem as regiões ativas, e a fração de energia perdida por ionização gera o sinal do detector. Os elétrons resultantes da ionização derivam sob o campo elétrico aplicado à fenda de LAr, produzindo uma corrente proporcional à energia depositada, que é então amplificada e moldada por um filtro bipolar do tipo CR–RC^2, amostrada a uma taxa de 40 MHz (a cada 25 ns) e digitalizada pela cadeia de leitura (COLLABORATION, 2017; ATLAS Collaboration, 2019).

Em cada célula de detecção, o formato do pulso é aproximadamente triangular como pode ser visto na Figura 12

Figura 12 – Forma do pulso de corrente do ECAL e do sinal de saída formado e amostrado. Os pontos indicam uma posição ideal das amostras separadas por 25 ns.



Fonte: (ATLAS Collaboration, 2019)

.

A energia depositada em cada célula é estimada a partir da amplitude do sinal, AA, convertida em contagens pelo conversor analógico-digital (ADC), conforme a Eq. (2.2).

$$A = \sum_{j=1}^{N_{amostras}} a_j(s_j - p)., \tag{2.2}$$

em que p é o pedestal (base estimada para os pulsos produzidos pelos sensores do calorímetro), s_j são as amostras do pulso produzido pelo calorímetro, a_j são os pesos obtidos a partir do filtro ótimo projetado para o canal de interesse, e N_{amostras} representa o número de amostras utilizadas no cálculo da energia E. A expressão para a energia estimada E é:

$$E = F_{\mu A} \rightarrow \text{MeV} \times F_{DAC} \rightarrow \mu A \times \frac{1}{\frac{M_{phys}}{M_{cali}}} \times G \times A; \tag{2.3}$$

Na Eq. (2.3) temos a energia estimada em uma célula do ECAL, onde:

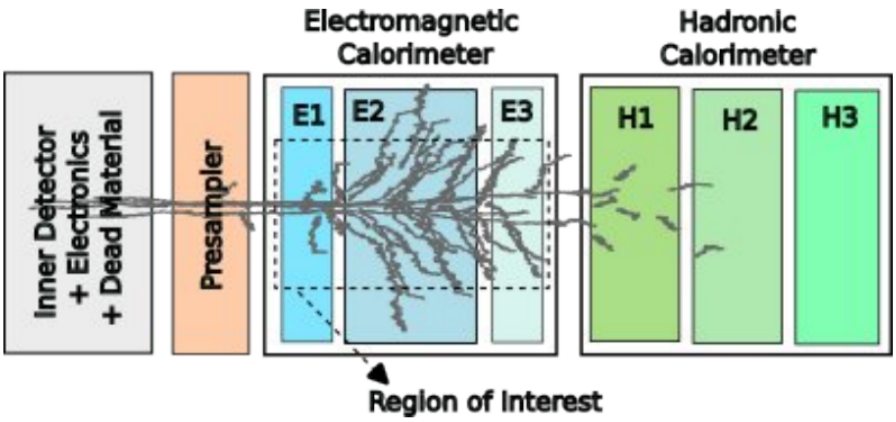
- $F_{\mu A} \rightarrow$ MeV é um fator que relaciona a corrente de ionização no calorímetro com a energia depositada. Ele depende de fatores como a razão de amostragem do calorímetro. Seu valor é determinado a partir de procedimentos experimentais de calibração (ATLAS Collaboration, 2019).
- $F_{ADC} \rightarrow \mu$A é o fator de conversão de contagens ADC em corrente de entrada. Ele é obtido a partir de execuções dedicadas de calibração da eletrônica.
- $\frac{M_{phys}}{M_{cali}}$ quantifica a relação entre os máximos dos pulsos físicos e de calibração correspondentes à mesma corrente de entrada.

- G é o ganho computado injetando um sinal calibrado conhecido e reconstruindo a resposta correspondente da célula.
- A é a amplitude estimada do sinal digitalizado em contagens ADC.

A partir da energia estimada em cada célula pela Eq. (2.3), calcula-se a energia de um candidato a partícula eletromagnética (elétron, pósitron ou fóton) por meio da soma das energias das células localizadas na região de interesse (RoI) do calorímetro eletromagnético, destacada na Figura 13.

As principais fontes de incerteza na estimação de energia derivam de três fatores críticos: (i) perda de informação lateral causada pelo vazamento da cascata eletromagnética além da região de interesse (RoI); (ii) perdas longitudinais nas camadas hadrônicas do calorímetro; e (iii) interações da partícula com material do detector antes de sua entrada no calorímetro, fenômeno conhecido como perdas *upstream*. Estes efeitos podem levar à subestimação ou superestimação da energia depositada, conforme ilustrado na Figura 13.

Figura 13 – Ilustração das principais causas dos erros na estimação da energia total da partícula ao interagir com o calorímetro do ATLAS: perda de energia antes (*upstream*) de interagir com o detector; vazamento lateral e longitudinal além da região de interesse.



Fonte: (SIMAS FILHO et al., 2021)

A etapa de calibração é importante para garantir que as medições de energia realizadas pelo calorímetro estejam precisas e confiáveis. A calibração inadequada pode comprometer a qualidade da seleção de candidatos a elétrons.

Neste capítulo, foram apresentados os conceitos fundamentais da física de partículas e do experimento ATLAS, que fornecem o contexto científico para o desenvolvimento deste trabalho. O capítulo seguinte será dedicado à fundamentação teórica das técnicas de aprendizado de máquina utilizadas na metodologia desta pesquisa.

3 Aprendizado de Maquina e Árvores de Decisão

3.1 Aprendizado de Máquina

Um algoritmo de aprendizado de máquina é um algoritmo que consegue aprender através dos dados (GOODFELLOW; BENGIO; COURVILLE, 2016). Para Tom M. Mitchell, a definição de aprendizado de máquina é: "Um programa é dito aprender pela experiência **E** com respeito a uma classe de tarefas **T** e desempenho mensurado por **P**, se o desempenho das tarefas **T**, mensurado por **P**, melhora com a experiência **E**."

Tipicamente, tarefas de aprendizado de máquina são descritas em termos do processamento de uma coleção de atributos de um objeto ou evento (GOODFELLOW; BENGIO; COURVILLE, 2016). Usualmente, representa-se um exemplo como um vetor:

$$\mathbf{x} \in \mathbb{R}^D,$$

onde cada dimensão x_i do vetor é um atributo: $\mathbf{x^T} = [x_1, x_2, ..., x_D]$.

Há diversos tipos de tarefas que podem ser resolvidas com aprendizado de máquina, nesse trabalho focaremos na regressão que será discutida nas seções seguintes.

3.1.1 Regressão

O objetivo da regressão é fazer predições do valor de uma ou mais variáveis contínuas y, dado um vetor de D dimensões de variáveis de entrada $\mathbf{x}$ (BISHOP, 2006).

Sendo assim, para um conjunto de dados de treinamento composto por N observações $\{\mathbf{x}_n\}$, onde $n = 1, \ldots, N$, juntamente com os valores alvo correspondentes $\{y_n\}$, o objetivo da regressão é estimar o valor de y associado a um novo valor de $\mathbf{x}$. Da forma mais simples, isso pode ser feito ajustando um modelo apropriado $\hat{y}(\mathbf{x})$, cujas previsões correspondem aos valores estimados de y para novas entradas $\mathbf{x}$ (BISHOP, 2006).

A forma mais simples do modelo de regressão é um modelo linear, que envolve uma combinação linear das variáveis de entrada:

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \cdots + w_D x_D, \tag{3.1}$$

na qual $\mathbf{x} = [x_1, \ldots, x_D]^T$ e $\mathbf{w} = [w_1, \ldots, w_D]^T$. Este modelo é conhecido como regressão linear. A principal característica deste modelo é que ele é uma função linear dos parâmetros $w_0, \ldots, w_D$. No entanto, também é uma função linear das variáveis de entrada x_i, o que impõe limitações significativas ao modelo.

De forma geral, sob uma perspectiva probabilística, buscamos modelar a distribuição preditiva condicional $p(y \mid \mathbf{x})$, que expressa nossa incerteza sobre o valor de y dado $\mathbf{x}$ (BISHOP, 2006). A

escolha da forma de $p(y \mid \mathbf{x})$ depende do tipo de variável de resposta: para variáveis contínuas, é comum assumir uma distribuição Gaussiana; para variáveis categóricas, utilizam-se distribuições Bernoulli ou multinomial.

Para capturar relações não lineares, introduzimos um conjunto de funções base $\{\psi_j(\mathbf{x})\}_{j=0}^{D}$, em que $\psi_0(\mathbf{x}) = 1$ garante o termo de viés. Definindo o vetor de características transformadas

$$\boldsymbol{\psi}(\mathbf{x}) = [\psi_0(\mathbf{x}), \psi_1(\mathbf{x}), \ldots, \psi_D(\mathbf{x})]^{\mathsf{T}},$$

o modelo estendido fica:

$$\hat{y}(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{D} w_j \psi_j(\mathbf{x}), \tag{3.2}$$

na equação anterior, temos que o total de parâmetros desse modelo será $D + 1$ (BISHOP, 2006).

O parâmetro w_0 permite qualquer deslocamento fixo nos dados e é chamado de parâmetro de viés. É conveniente definir uma função base adicional $\psi_0(\mathbf{x}) = 1$, dessa forma, temos:

$$\hat{y}(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{D} w_j \psi_j(\mathbf{x}) = \mathbf{w}^T \psi(\mathbf{x}), \tag{3.3}$$

na Eq. (3.3), os vetores são definidos como

$$\mathbf{w} = [w_0, \ldots, w_D]^{\mathsf{T}} \quad \text{e} \quad \boldsymbol{\psi}(\mathbf{x}) = [\psi_0(\mathbf{x}), \ldots, \psi_D(\mathbf{x})]^{\mathsf{T}}.$$

Portanto, em problemas de regressão, deve-se encontrar os multiplicadores $\mathbf{w}$ de modo que a previsão $\hat{y}$ seja mais próxima possível da saída real y. Para essa finalidade, focaremos no método do gradiente descendente.

3.1.1.1 Gradiente Descendente

O algoritmo do gradiente descendente é um dos algoritmos de otimização mais comuns e uma das formas mais simples de resolver o problema de encontrar os valores ótimos de $\mathbf{w}$. Define-se uma medida ou função de custo $L(\mathbf{w})$ para inferir se o modelo está próximo do desempenho desejado e encontra-se os valores de $\mathbf{w}$ a fim de minimizar essa função. Alguns exemplos de funções de custo são:

- Erro médio quadrático ou *Mean Squared Error* (MSE):

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{3.4}$$

- Erro médio absoluto ou L1 ou *Mean Absolute Error* (MAE):

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{3.5}$$

- Erro Médio Absoluto Percentual ou *Mean Absolute Percentage Error* (MAPE):

$$\text{MAPE} = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{3.6}$$

Nas Equações 3.4, 3.5, 3.6, apresentadas a seguir, m é o número de observações ou amostras no conjunto de dados. Dada a função de custo define-se uma direção de busca de $\mathbf{w}$ conforme o gradiente da função de perda. No Algoritmo 1 está um exemplo de sua utilização com a função de custo MSE. Há também algoritmos modificados da descida por gradiente que levam em conta as características específicas dos dados e ajustam as taxas de aprendizado de maneira mais adaptativa. Alguns exemplos incluem o RMSProp e o AdaGrad (HAYKIN, 2008).

Algoritmo 1 Descida de Gradiente para Regressão Linear

Entrada: Características $\mathbf{X}$, Alvo $\mathbf{y}$, taxa de aprendizado α, número de iterações n_{iter}
Saída: Pesos $\mathbf{w}$

1: Inicialize os pesos $\mathbf{w}$ com pequenos valores aleatórios
2: **para** $i = 1$ até n_{iter} **do**
3: Calcule as previsões $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$
4: Calcule o custo $L(\mathbf{w}) = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2$, onde m é o número de amostras
5: Calcule o gradiente $\nabla L = \frac{1}{m}\mathbf{X}^T(\hat{\mathbf{y}} - \mathbf{y})$
6: Atualize os pesos $\mathbf{w} \leftarrow \mathbf{w} - \alpha\nabla L$
7: **end para**

3.1.2 Classificação

Diferente da regressão, há casos em que a variável y não é quantitativa, mas sim qualitativa. Por exemplo, a cor de um carro: nesse caso, chamamos essa variável de categórica, e as tarefas que visam predizer uma categoria com base em atributos de entrada são denominadas classificação. Métodos de classificação frequentemente envolvem a atribuição de probabilidade de uma observação pertencer a uma classe, seguida pela classificação propriamente dita, comportando-se, de certa forma, como modelos de regressão (JAMES et al., 2023).

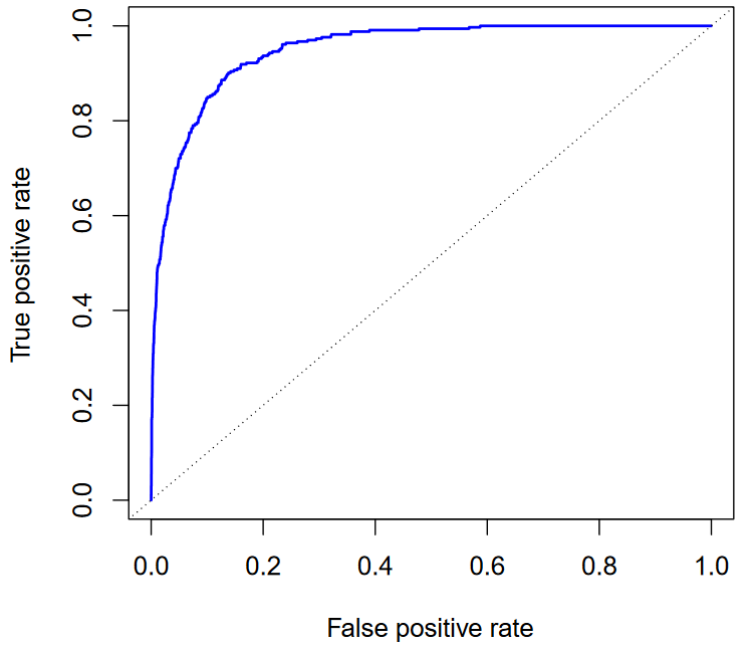
A classificação pode ser binária (isto é, quando o modelo distingue entre duas classes distintas) ou multiclasse (com múltiplas categorias de saída). Técnicas de aprendizado de máquina têm sido amplamente utilizadas para resolver problemas dessa natureza em diversos campos, como na área médica para classificação de melanomas (KAUR et al., 2022) ou na identificação de defeitos em superfícies de chapas metálicas (LUO et al., 2023).

Para classificação binária, é possível definir medidas de desempenho que quantificam a eficácia do modelo. Seja y_i a classe verdadeira da instância i e $\hat{y}_i$ a classe predita pelo modelo, onde $y_i, \hat{y}_i \in \{0, 1\}$. Definimos a probabilidade de detecção (P_D) como a probabilidade de o modelo classificar corretamente uma amostra positiva, e a probabilidade de falso alarme (P_F) como a probabilidade de classificar erroneamente uma amostra negativa como positiva.

Uma forma de avaliar o desempenho de um classificador binário é por meio da curva *Receiver Operator Characteristic* (ROC) (MARZBAN, 2004). Essa curva relaciona P_D (no eixo das ordenadas) e P_F (no eixo das abscissas) para diferentes limiares de decisão, assumindo que o modelo retorna probabilidades de pertencimento a uma classe. A área sob a curva ROC (*Area*

Under the Curve (AUC)) fornece uma medida agregada de desempenho, atingindo valor máximo (1,0) quando o classificador é perfeito. Na Figura 14, por exemplo, a AUC é 0,95, indicando excelente capacidade de discriminação por estar próxima do valor ideal (JAMES et al., 2023).

Figura 14 – Exemplo de Curva ROC



Fonte: (JAMES et al., 2023)

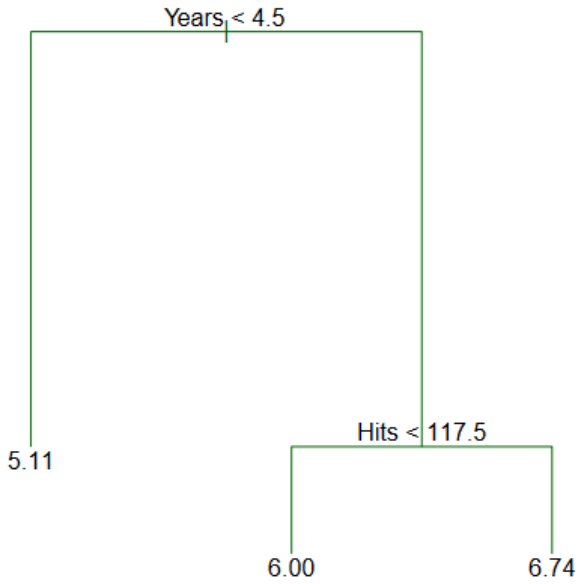
3.2 Árvores de Decisão para Regressão

As árvores de decisão são modelos de aprendizado de máquina supervisionados amplamente empregados tanto para tarefas de classificação quanto de regressão. Atualmente é utilizada para calibração de energia na etapa de reconstrução *offline* de elétrons e fótons do experimento ATLAS (ATLAS Collaboration, 2019).

Uma árvore de decisão é um modelo com uma estrutura hierárquica semelhante a um fluxograma. Nessa estrutura, cada nó interno, também chamado de nó de decisão, avalia uma condição sobre um atributo dos dados. Cada ramo representa o resultado dessa avaliação e direciona o fluxo para um nó subsequente. Os nós que finalizam os caminhos da árvore são os nós folha, ou nós terminais, e contêm o valor previsto para a variável alvo.

Um exemplo dessa estrutura é ilustrado na Figura 15, os atributos usados como exemplo são *Years* e *Hits*. Os nós de decisão aplicam condições a esses atributos, como $Years < 4.5$ e $Hits < 117.5$. Ao final de cada caminho, os nós folha indicam o valor atribuído a um exemplo que satisfaça todas as condições daquele percurso.

Figura 15 – Exemplo de Árvore de Decisão para Regressão



Fonte: (HASTIE; TIBSHIRANI; FRIEDMAN, 2009)

Definido um conjunto de dados de entrada $\mathbf{X}$, representado por uma matriz onde cada linha é $\mathbf{x}_i = [x_{i1}, \ldots, x_{ip}]$ corresponde às p variáveis de entrada para a i-ésima observação. O algoritmo precisa decidir como particionar as variáveis e seus pontos de corte, e também a topologia ou formato da árvore (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Suponha uma partição em M regiões $R_1, R_2, \ldots, R_M$, onde R_m representa uma partição qualquer dentre as M regiões, e modela-se a resposta como uma constante γ_m em cada região, conforme a seguinte equação:

$$f(x) = \sum_{m=1}^{M} \gamma_m I(\mathbf{x} \in R_m), \tag{3.7}$$

onde, $I(\mathbf{x} \in R_m)$, é definida como:

$$I(\mathbf{x} \in R_m) = \begin{cases} 1, & \text{se } \mathbf{x} \in R_m \\ 0, & \text{caso contrário.} \end{cases} \tag{3.8}$$

Caso seja adotada como função de perda a *Residual Sum of Squares* (RSS), ou soma dos quadrados, $\sum(y_i - f(\mathbf{x_i}))^2$ é possível demonstrar que o melhor valor de γ_m, denotado por $\hat{\gamma}_m$, é a média de y_i na região R_m (HASTIE; TIBSHIRANI; FRIEDMAN, 2009), isto é:

$$\hat{\gamma}_m = \text{média}(y_i \mid \mathbf{x_i} \in R_m). \tag{3.9}$$

Entretanto, encontrar o melhor particionamento considerando todas as possíveis partições do espaço de características J é computacionalmente inviável (JAMES et al., 2023). Por essa razão adota-se uma abordagem gananciosa de cima para baixo, conhecida como divisão binária recursiva. Começando com todo o conjunto de dados, considera-se uma variável de corte j e

uma variável de corte t, define-se um par de semi-planos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009):

$$R_1(j,t) = \mathbf{X} \mid \mathbf{X}_j \leq t \text{ e } R_2(j,t) = \mathbf{X} \mid \mathbf{X}_j > t. \tag{3.10}$$

Então, procura-se a variável j e o ponto de corte t que resolvem o problema de otimização:

$$\min_{j,t} \left[\min_{\gamma_1} \sum_{\mathbf{x}_i \in R_1(j,t)} (y_i - \gamma_1)^2 + \min_{\gamma_2} \sum_{\mathbf{x}_i \in R_2(j,t)} (y_i - \gamma_2)^2 \right], \tag{3.11}$$

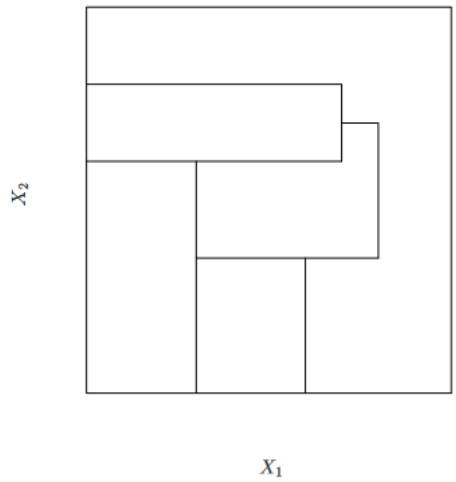
para qualquer escolha de j e t, o problema de otimização interno é resolvido por:

$$\hat{\gamma}_1 = \text{média}(y_i \mid \mathbf{x_i} \in R_1(j,t) \text{ e } \hat{\gamma}_2 = \text{média}(y_i \mid \mathbf{x_i} \in R_2(j,t)). \tag{3.12}$$

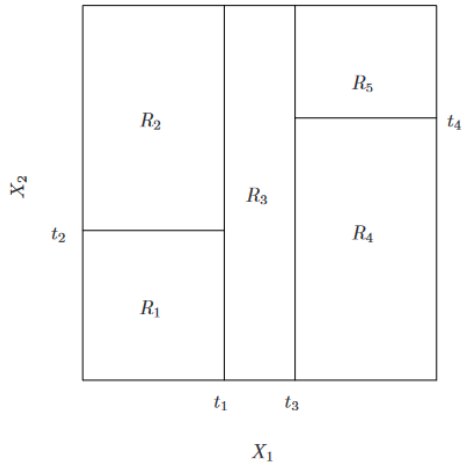
Para cada variável de divisão, a determinação do ponto de divisão t pode ser feita muito rapidamente e, portanto, ao percorrer todas as entradas, a determinação do melhor par (j,t) é viável. Após encontrar a melhor divisão, particionamos os dados nas duas regiões resultantes e repete-se o processo de divisão em cada uma das duas regiões. Em seguida, esse processo é repetido em todas as regiões resultantes (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

A Figura 16 apresenta uma visualização completa do funcionamento de uma árvore de decisão em um espaço de características bidimensional. A Figura 16(a) serve como um contraexemplo, mostrando uma partição com fronteiras complexas que não poderiam ser geradas por uma árvore, a qual utiliza apenas divisões binárias e recursivas. Em contraste, a Figura 16(b) ilustra o particionamento real produzido pelo algoritmo, resultando em cinco regiões retangulares distintas. A estrutura hierárquica que gera essa divisão é exibida na Figura 16(c), onde os cinco nós folha (regiões $R_1, \ldots, R_5$) correspondem diretamente às cinco áreas. Por fim, a Figura 16(d) traduz essa estrutura em uma superfície de predição tridimensional, na qual o valor previsto é constante dentro de cada uma das cinco regiões definidas pela árvore.

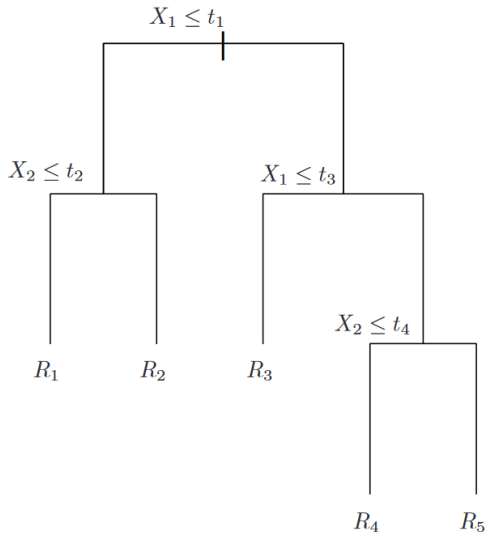
Figura 16 – Visualizações das partições do espaço de características bidimensional e da correspondente árvore de decisão, junto com um gráfico de perspectiva da superfície de previsão.



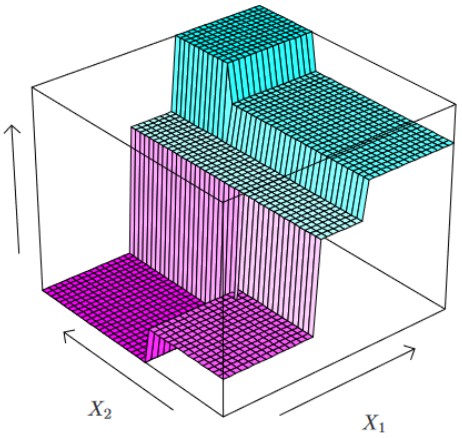
(a) Partição do espaço bidimensional que não resulta da divisão binária recursiva.



(b) Saída da divisão binária recursiva em um exemplo bidimensional.



(c) Árvore correspondente à partição na figura (b).



(d) Gráfico de perspectiva da superfície de previsão correspondente à árvore na figura (c).

Fonte: (HASTIE; TIBSHIRANI; FRIEDMAN, 2009)

Esse processo pode produzir boas predições no conjunto de treinamento, mas é muito suscetível ao sobreajuste levando a um baixo desempenho no conjunto de teste, isso ocorre pois a árvore resultante pode ser muito complexa. Para lidar com esse problema, uma solução comum é o processo de poda (do inglês, *tree pruning*), que consiste em reduzir a complexidade da árvore para evitar o sobreajuste (*overfitting*) e melhorar sua capacidade de generalização. Neste trabalho, será abordada especificamente a técnica de pós-poda, na qual a árvore é primeiro construída por completo para depois ser simplificada.

A ideia é construir uma árvore complexa T_0 e então podá-la para obter uma subárvore, $T \subset T_0$, que obtenha o menor erro no conjunto de teste. Para encontrar T_0, define-se o critério

de complexidade de custo $C_\lambda(T)$ (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Seja uma *subárvore* qualquer árvore que possa ser obtida a partir de T_0 por meio da poda, isto é, colapsando-se qualquer número de seus nós internos. Indexamos os nós terminais por m, sendo que cada nó m representa uma região R_m. Seja $|T|$ o número de nós terminais da árvore T; então:

$$N_m = \#\mathbf{x}_i \in R_m, \tag{3.13}$$

$$\hat{\gamma}_m = \frac{1}{N_m} \sum_{\mathbf{x}_i \in R_m} y_i, \tag{3.14}$$

$$Q_m(T) = \frac{1}{N_m} \sum_{\mathbf{x}_i \in R_m} (y_i - \hat{\gamma}_m)^2, \tag{3.15}$$

em que N_m é o número de observações $\mathbf{x}_i$ na região R_m (neste contexto, o símbolo $\#$ é usado para representar a contagem de elementos), e $Q_m(T)$ é a medida de impureza do nó, correspondente ao erro quadrático médio. Define-se então o critério de custo de complexidade $C_\lambda(T)$:

$$C_\lambda(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \lambda \mid T \mid . \tag{3.16}$$

E tem-se como objetivo encontrar, para cada λ, a subárvore $T_\lambda \subseteq T$ que minimize $C_\lambda(T)$. O parâmetro λ governa a troca entre o tamanho da árvore e o quão bem ela se encaixa no conjunto de dados, ou, o compromisso entre viés e variância. Maiores valores de λ resultam em árvores menores e para $\lambda = 0$ temos como solução a árvore original.

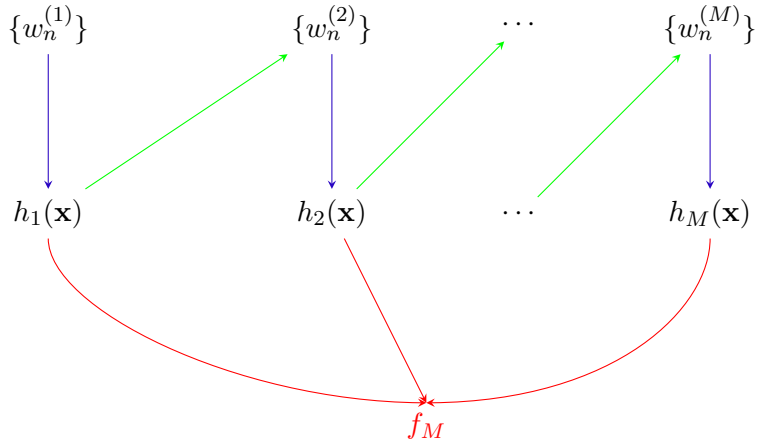
Para cada λ, é possível mostrar que existe uma única subárvore menor T_λ que minimiza $C_\lambda(T)$. Para encontrar T_λ, utiliza-se a poda do elo mais fraco: sucessivamente, colapsa-se o nó interno que produz o menor aumento por nó em $\sum_{m=1}^{M} N_m Q_m(T)$ e continua-se até chegar na raiz da árvore. Isso resulta em uma sequência (finita) de subárvores, e pode-se mostrar que essa sequência deve conter T_λ (RIPLEY, 1996). A estimativa de λ é alcançada por meio de validação cruzada[1] de cinco ou dez vezes: escolhe-se o valor de λ para minimizar RSS validados cruzadamente.

3.2.1 *Gradient Boosting*

O *Gradient Boosting Decision Trees* (GBDT) é um método proposto por (FRIEDMAN, 2001), é baseado na ideia dos algoritmos de *ensemble*, em particular o *boosting*, que tem como objetivo combinar classificadores fracos para formar um comitê forte que produza boas predições (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). A Figura 17 apresenta uma representação esquemática desse modelo.

[1] Validação cruzada é uma técnica de avaliação de modelos que consiste em dividir o conjunto de dados em várias partes, utilizando algumas delas para treinar o modelo e outras para testar, de forma repetida, para obter uma estimativa mais robusta da performance do modelo.

Figura 17 – Ilustração esquemática do algoritmo *boosting*. Cada regressor base $h_m(\mathbf{x})$ é treinado em uma versão ponderada do conjunto de treinamento (setas azuis), na qual os pesos $w_n^{(m)}$ são ajustados de acordo com o desempenho do regressor anterior $h_{m-1}(\mathbf{x})$ (setas verdes). Depois que todos os regressores base forem obtidos, eles são somados para formar o modelo final $f_M(\mathbf{x})$, como indicado pelas setas vermelhas.



Fonte: Autoria própria, baseado em (BISHOP, 2006).

Como visto na seção anterior uma árvore de regressão particiona o espaço das variáveis preditoras em regiões distintas $R_j,\ j = 1, 2, \ldots, J$ que representam os nós terminais da árvore. E uma constante γ_j é definida para cada região e a regra de predição é:

$$\mathbf{x} \in R_j \implies f(\mathbf{x}) = \gamma_j. \tag{3.17}$$

De tal forma, uma árvore pode ser formalmente expressa por:

$$T(\mathbf{x}; \Theta) = \sum_{j=1}^{J} \gamma_j I(\mathbf{x} \in R_j), \tag{3.18}$$

com parâmetros $\Theta = \{R_j, \gamma_j\}_1^J$. Os parâmetros da árvore podem ser encontrados minimizando o a função de custo L:

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{j=1}^{J} \sum_{\mathbf{x}_i \in R_j} L(y_i, \gamma_j). \tag{3.19}$$

Há diversas formas de encontrar uma solução para a Eq. (3.19), geralmente sub-ótimas devido a complexidade do problema (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). O modelo *boosted* é a soma destas árvores Eq. (3.20), onde M é o número de árvores do comitê. Induzido de maneira progressiva, i.e, a cada passo deve-se resolver a Eq. (3.21), para o conjunto de regiões e constantes $\Theta_m = \{R_{jm}, \gamma_{jm}\}_1^{J_m}$ da próxima árvore dado o modelo presente $f_{m-1}(x)$, onde N é o número de amostras do conjunto de treinamento.

$$f_M(x) = \sum_{m=1}^{M} T(\mathbf{x}; \Theta_m), \tag{3.20}$$

$$\hat{\Theta} = \arg\min_{\Theta_m} \sum_{i=1}^{N} L(y_i, f_{m-1}(\mathbf{x}_i) + T(\mathbf{x}_i; \Theta_m)). \tag{3.21}$$

A solução da Eq. (3.21) pode ser obtida por algoritmos numéricos, desde que a função de perda L seja diferenciável. Seja $L(f)$ a função que mede o erro de $f(\mathbf{x})$, preditor de y no conjunto de treinamento, conforme Eq. (3.22).

$$L(f) = \sum_{i=1}^{N} L(y_i, f(\mathbf{x}_i)). \tag{3.22}$$

O objetivo é minimizar $L(f)$ com respeito a f, onde $f(x)$ tem a restrição de ser uma soma de árvores (3.18). Ignorando essa restrição, pode-se tomar esse problema como uma otimização numérica:

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} L(\mathbf{f}), \tag{3.23}$$

na qual $\mathbf{f} \in \mathbb{R}^N$ são os valores aproximados da função $f(\mathbf{x}_i)$ a cada um dos N pontos $\mathbf{x}_i$, $\mathbf{f} = \{\mathbf{f}(\mathbf{x}_1), \mathbf{f}(\mathbf{x}_2), \ldots, \mathbf{f}(\mathbf{x}_N)\}^T$. A otimização numérica resolve a Eq. (3.23) como uma soma de componentes de vetores da Eq. (3.24), onde $\mathbf{f}_0 = \mathbf{h}_0$ é uma estimativa inicial e a cada passo $\mathbf{f}_m$ é calculada baseada no parâmetro $\mathbf{f}_{m-1}$ (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

$$\mathbf{f}_M = \sum_{m=0}^{M} \mathbf{h}_m, \ \mathbf{h}_m \in \mathbb{R}^N. \tag{3.24}$$

Como discutido na seção 3.1.1.1, pode-se resolver problemas deste tipo utilizando o gradiente descendente. O passo de atualização é dado por $\mathbf{h}_m = -\rho_m \mathbf{g}_m$, onde ρ_m é um escalar e $\mathbf{g}_m$ denota o gradiente de $L(\mathbf{f})$ avaliado em $\mathbf{f} = \mathbf{f}_{m-1}$. Os componentes do vetor gradiente $\mathbf{g}_m$ são dados por:

$$g_{im} = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x_i)=fm-1(x_i)}, \tag{3.25}$$

a cada passo ρ_m é dado como a solução para a Eq. (3.26):

$$\rho_m = \arg\min_{\rho} L(\mathbf{f}_{m-1} - \rho \mathbf{g}_m). \tag{3.26}$$

A solução é portanto atualizada para:

$$\mathbf{f_m} = \mathbf{f_{m-1}} - \rho_\mathbf{m} \mathbf{g_m}, \tag{3.27}$$

e o processo é repetido na próxima iteração. O método do gradiente descendente pode ser visto como uma estratégia gananciosa (do inglês, *greedy*), já que $-\mathbf{g}_m$ é a direção de descida local em $\mathbb{R}^N$ na qual a função de perda $L(f)$ decresce mais rapidamente no ponto $f = \mathbf{f}_{m-1}$ (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

No *gradient boosting* uma árvore $T(\mathbf{x}; \Theta_m)$ é induzida de maneira progressiva, onde em cada etapa a árvore solução é aquela que reduz ao máximo a Eq. (3.21), dada a modelagem atual f_{m-1} e seus ajustes $f_{m-1}(\mathbf{x}_i)$. Assim, as predições da árvore $T(\mathbf{x}_i; \Theta_m)$ são análogas aos componentes do gradiente negativo. A principal diferença entre eles é que os componentes da árvore $\mathbf{t}_m = (T(\mathbf{x}_1; \Theta_m), \ldots, T(\mathbf{x}_N; \Theta_m))$ não são independentes, eles são as previsões de uma árvore de decisão com J_m nós terminais, enquanto o gradiente negativo é a direção de descida máxima não restrita.

Se minimizar a perda nos dados de treinamento fosse o único objetivo, o gradiente descendente seria a estratégia preferida. O gradiente é trivial de calcular para qualquer função de perda diferenciável $L(y, f(\mathbf{x}))$. Infelizmente, o gradiente é definido apenas nos pontos de dados de treinamento, enquanto o objetivo final é generalizar $f_M(\mathbf{x})$ para novos dados não representados no conjunto de treinamento (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Uma possível solução para esse problema é induzir uma árvore $T(\mathbf{x}; \Theta_m)$ na m-ésima iteração cujas previsões t_m sejam as mais próximas possíveis do gradiente negativo. Utilizando o erro quadrático para medir a proximidade, temos:

$$\tilde{\Theta}_m = \arg\min_{\Theta} \sum_{i=1}^{N} (-g_{im} - T(\mathbf{x}_i; \Theta))^2 \tag{3.28}$$

Ou seja, ajusta-se a árvore T aos valores do gradiente negativo por mínimos quadrados. Embora as regiões de solução $\tilde{R}_{jm}$ não sejam idênticas às regiões R_{jm} que resolvem 3.21, geralmente são suficientemente semelhantes para servir ao mesmo propósito. Em qualquer caso, o procedimento de indução progressiva e indução de árvore de decisão de cima para baixo são procedimentos de aproximação (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). O algoritmo GBDT pode ser visto em 2.

Algoritmo 2 Gradient Tree Boosting

Entrada: Inicialize $f_0(x) = \arg\min_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$
Resultado: $\hat{f}(x) = f_M(x)$

para $m = 1$ to M **do**
  para $i = 1$ to N **do**
    Calcule $r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}$
  end para
  Ajuste uma árvore de regressão aos alvos r_{im}, fornecendo regiões terminais R_{jm}, $j = 1, 2, ..., J_m$
  para $j = 1$ to J_m **do**
    Calcule $\gamma_{jm} = \arg\min_\gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$
  end para
  Atualize $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$
end para

É necessário pontuar, que é importante adotar estratégias para evitar o sobreajuste ao utilizar o algoritmo 2. Essas estratégias podem incluir técnicas de regularização, como a poda de árvores,

e a validação cruzada para seleção de hiperparâmetros. Detalhes sobre essas estratégias podem ser encontrados em obras como (JAMES et al., 2023) e (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

3.2.1.1 *Light GBM*

O *Light Gradient Boosting Machine* (LGBM), proposto pela Microsoft em 2017 (KE et al., 2017), surge como uma evolução eficiente do paradigma GBDT para enfrentar desafios de escalabilidade em grandes volumes de dados. Enquanto o GBDT tradicional enfrenta limitações computacionais devido ao crescimento geométrico do custo com dados massivos, o LightGBM mantém a precisão do modelo enquanto otimiza e a velocidade de treinamento e o consumo de memória (JU et al., 2019).

A principal inovação do LightGBM reside em duas estratégias: um *método baseado em histogramas* e um *crescimento folha a folha* (*leaf-wise*). O método baseado em histogramas substitui a busca exaustiva por pontos de divisão contínuos, agrupando valores em intervalos discretos (*bins*). Essa abordagem reduz significativamente o custo computacional, pois a construção dos histogramas tem custo $O(\#\text{bins} \times \#\text{atributos})$, enquanto a busca pelos pontos de divisão tem custo $O(\#\text{dados} \times \#\text{atributos})$. Como o número de *bins* é geralmente muito menor que o número de dados, a construção dos histogramas tende a dominar a complexidade computacional, viabilizando o processamento de grandes *datasets* com eficiência espacial. Adicionalmente, o processo de *binning* atua como regularização implícita, mitigando o sobreajuste.

Na construção das árvores, o LightGBM adota uma estratégia assimétrica de crescimento *leaf-wise*, em contraste ao método tradicional *level-wise* (crescimento por níveis). Em vez de expandir todas as folhas simultaneamente em cada nível, o algoritmo seleciona dinamicamente a folha com maior redução de perda para divisão (Figura 18). Isso permite maior capacidade de modelagem com a mesma profundidade máxima, embora requeira ajuste cuidadoso de parâmetros como *max_depth* para evitar sobreajuste.

A combinação dessas técnicas permite ao LGBM superar algumas limitações do GBDT, especialmente em cenários com dados esparsos ou de alta dimensionalidade. Sua implementação paralelizável e suporte nativo a *early stopping* consolidaram-no como referência em competições de *machine learning* e aplicações industriais onde eficiência e precisão são críticas (BENTÉJAC; CSÖRGŐ; MARTÍNEZ-MUÑOZ, 2020).

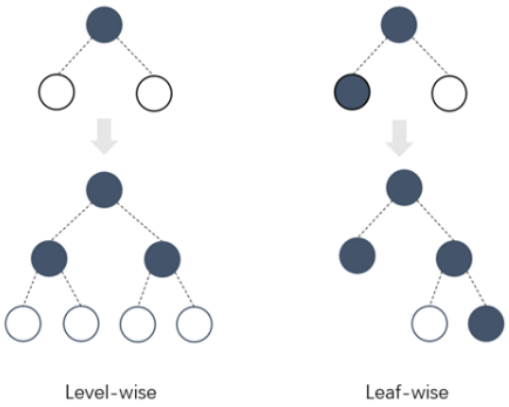


Figura 18 – Comparação entre o crescimento de árvores *Level-wise* e *Leaf-wise* no LightGBM. O método *Level-wise* expande a árvore de forma uniforme, nível por nível, enquanto o método *Leaf-wise* expande a árvore de maneira desigual, focando em folhas que reduzem mais o erro, permitindo um crescimento mais profundo e rápido.

Fonte: (JU et al., 2019)

Na Tabela 3 estão listados os principais hiperparâmetros que devem ser ajustados em um GBDT, conforme a implementação da biblioteca *LightGBM*.

Tabela 3 – Principais hiperparâmetros do modelo LightGBM.

Hiperparâmetro	**Descrição**
`num_boost_round`	Número de iterações (*boosting rounds*) a serem realizadas.
`learning_rate`	Taxa de aprendizado que reduz a contribuição de cada árvore.
`early_stopping_rounds`	Número de iterações sem melhora para acionar a parada antecipada.
`num_leaves`	Número máximo de folhas em cada árvore.
`max_depth`	Profundidade máxima da árvore (pode limitar o número de folhas).
`feature_fraction`	Fração das características utilizadas na construção de cada árvore.
`bagging_fraction`	Fração dos dados amostrados para treinar cada árvore.
`bagging_freq`	Frequência (em iterações) para aplicar o *bagging*.
`lambda_l1`	Parâmetro de regularização L1.
`min_split_gain`	Ganho mínimo necessário para realizar uma divisão.

Fonte: LightGBM (KE et al., 2017).

Neste capítulo, foi apresentada a fundamentação teórica do aprendizado de máquina e, em particular, foram expostos os modelos de árvores de decisão e de árvores de decisão reforçadas por gradiente (GBDT), que constituem a base desta pesquisa. O capítulo seguinte será dedicado à metodologia desenvolvida.

4 Metodologia

No presente capítulo, serão detalhados os procedimentos adotados durante a pesquisa para atender aos objetivos propostos.

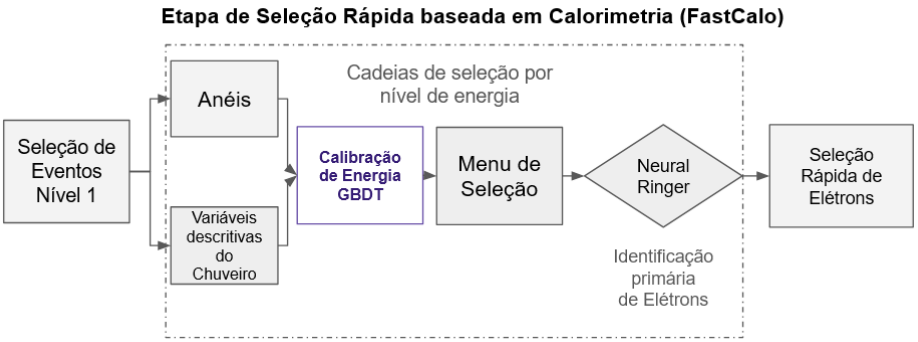
4.1 Calibração de Energia utilizando GBDT

Como discutido anteriormente, os processos de calibração são essenciais nas etapas do HLT e também na análise *offline* para a caracterização das assinaturas energéticas de interesse. Como apresentado, a energia estimada na etapa rápida pode ser descrita na Eq. (2.3). Neste trabalho, temos como objetivo adicionar um novo fator de calibração a ser estimado, denotado por α_{BDT}, que representa uma estimativa do valor de α definido por:

$$\alpha = \frac{E_T^{\text{Truth}}}{E_T^{\text{Fast}}}. \tag{4.1}$$

Como pode ser visto na Eq. (4.1), E_T^{Truth} e E_T^{Fast} são os valores de energia transversa verdadeira (obtidos por simulação de Monte Carlo) e estimada na etapa rápida, respectivamente. Na Figura 19 é possível ver o processo inserindo um novo bloco de processamento na etapa rápida para melhorar a eficiência da estimação de energia.

Figura 19 – Diagrama indicando o bloco de calibração proposto e sua integração com o HLT do ATLAS.

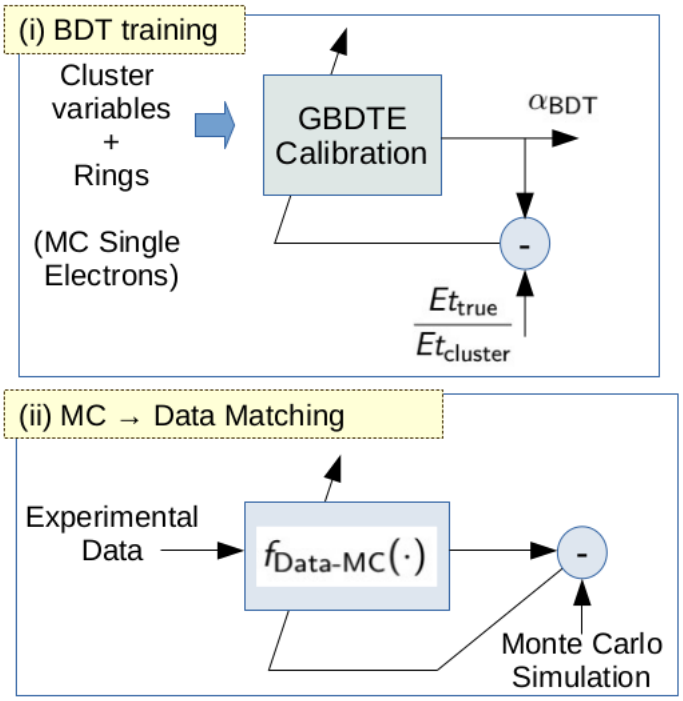


Fonte: Autoria Própria

O fator α_{BDT} será obtido a partir do modelo de aprendizado de máquina GBDT e inserido na parte rápida do HLT. Na Figura 19 é possível ver um diagrama das etapas de desenvolvimento do sistema, o valor de α_{BDT}. Na primeira etapa é feita a construção do modelo, são utilizadas as informações dos anéis e informações do agrupamento (*cluster*) de células como será apresentado nas próximas seções. Na segunda etapa é feito um ajuste entre as distribuições dos dados de simulação e experimentais para corrigir possíveis discrepâncias advindas das imperfeições na simulação. As informações de anéis já estão disponíveis no ambiente computacional da fase rápida do HLT, enquanto as variáveis de forma de chuveiro — normalmente empregadas em problemas

de calibração de energia — não são pré-computadas nesse estágio. Um dos objetivos é, portanto, avaliar e comparar duas estratégias de construção dos atributos de entrada para a GBDT: (i) exclusivamente a partir das características dos anéis concêntricos de células do calorímetro e (ii) a partir das variáveis tradicionais de forma de chuveiro.

Figura 20 – Diagrama das etapas de desenvolvimento do sistema de calibração.



Fonte: (SIMAS FILHO et al., 2021)

Portanto, a nova energia calibrada será dada por:

$$\text{Energia Estimada} \times \alpha_{BDT} = \text{Energia Calibrada}; \tag{4.2}$$

O procedimento de calibração, devido às características apresentadas anteriormente do detector, precisa ser projetado para diferentes intervalos de η e E_T. A partir de estudos prévios (ATLAS Collaboration, 2019), a colaboração ATLAS decidiu adotar a seguinte segmentação para o espaço de fases que também será adotada neste estudo:

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

Dessa forma, é necessário implementar um modelo para cada combinação de intervalos, totalizando $9 \times 13 = 117$ modelos. Além disso, todo o modelo foi incorporado ao ambiente computacional do ATLAS (Athena), a fim de avaliar seu desempenho no fluxo do *trigger* de elétrons com dados simulados no ambiente de produção do experimento.

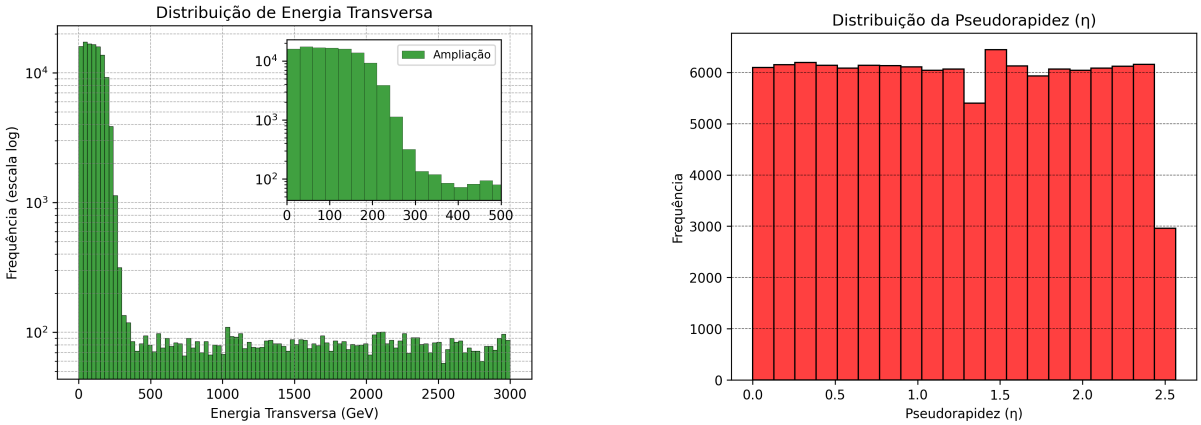
4.2 Dados Utilizados para o Treinamento

Os dados analisados neste estudo, para o treinamento e testes iniciais do modelo, foram cedidos pela Colaboração ATLAS e originam-se de simulações de eventos por meio do método de Monte Carlo para colisões que resultam na produção de elétrons isolados. Os dados derivados das simulações de Monte Carlo são aplicados no avanço da física de partículas, com o objetivo de modelar as operações do detector, tanto com os parâmetros atuais quanto em cenários futuros, foram utilizados aproximadamente 200.000 eventos.

A simulação também leva em conta o fenômeno de *pileup* que é denotado pela letra grega $<\mu>$. O fenômeno ocorre quando múltiplas colisões de prótons acontecem simultaneamente em um único evento de colisão. Isso resulta em uma sobreposição de sinais de diferentes eventos, tornando a identificação e caracterização dos eventos individuais mais desafiadora. Em outras palavras, é como se houvesse um "empilhamento" de eventos de colisão, dificultando a separação dos sinais de interesse do ruído de fundo (SOYEZ, 2019). É interessante haver muitas colisões pois elas aumentam a taxa de eventos observados, o que é essencial para a descoberta de fenômenos raros em colisores de partículas.

Na Figura 21 pode-se encontrar os histogramas característicos da energia transversa verdadeira e da pseudo-rapidez, η, para os eventos utilizados.

Figura 21 – Distribuições de E_T e $|\eta|$



(a) Histograma da Energia Transversa.

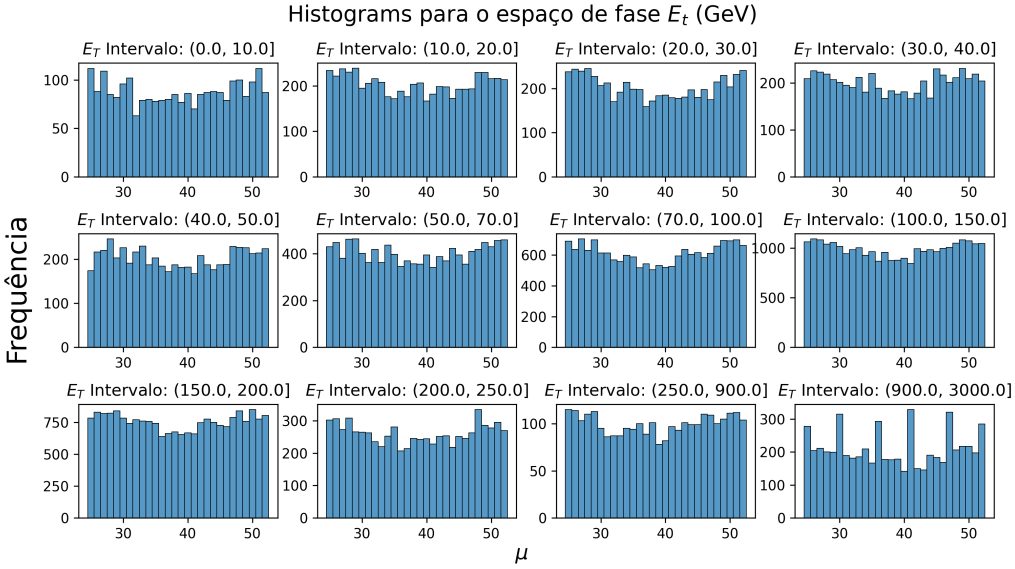
(b) Histograma da Pseudorapidez.

Fonte: Autoria Própria

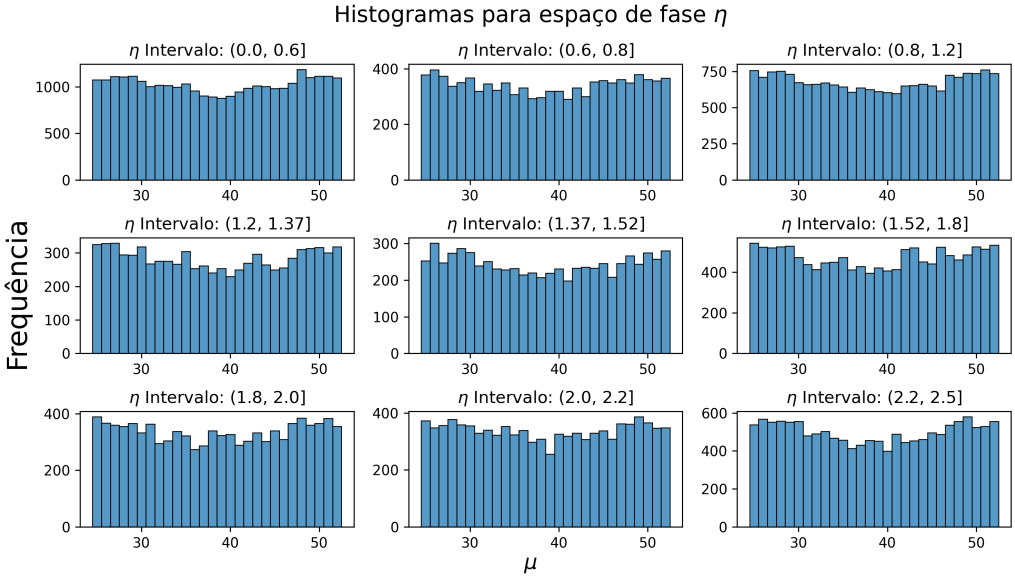
É possível perceber que a maioria dos eventos estão concentrados entre 0 e 300 GeV, porém estão distribuídos de modo aproximadamente uniforme em $|\eta|$. O perfil de deposição de energia do calorímetro muda em função de E_T, sendo que em baixas energias é mais difícil a identificação de elétrons. Também há uma variação em razão de $|\eta|$, o que ocorre devido às diferentes granularidades do sensores ao redor desse eixo, conforme apresentado anteriormente.

A distribuição de energia varia de 0 a 3000 GeV, enquanto a distribuição de η está entre 0 e 2,5. É importante também verificar o perfil de *pileups*, ou $< \mu >$, dos eventos. A Figura 22 mostra esse perfil, é notável destacar que a distribuição de *pileup* está distribuída aproximadamente uniformemente no intervalo de 25 a 52, dentro do espaço de fase.

Figura 22 – Distribuições de μ no espaço de fase E_T e $|\eta|$



(a) Histograma de μ no espaço de fase E_T.



(b) Histograma de μ no espaço de fase η.

Fonte: Autoria Própria

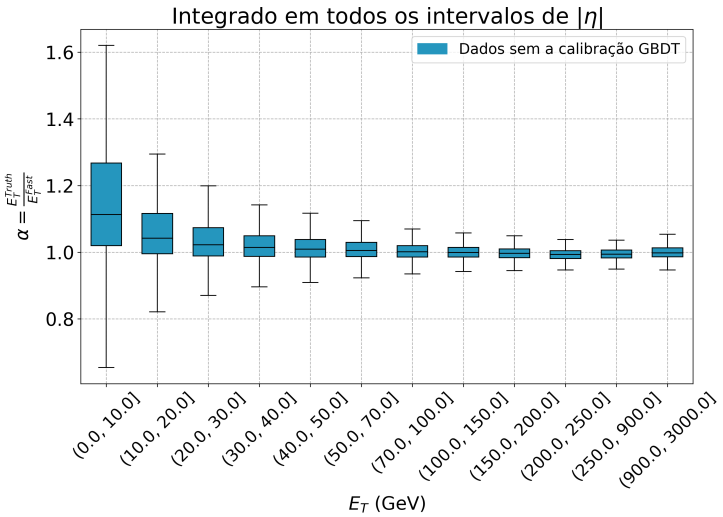
É possível verificar que μ varia entre 25 e 52, ou seja, ocorre essa quantidade de colisões simultâneas por evento. É importante destacar também que em todo o espaço de fase analisado a distribuição de μ é aproximadamente uniforme, tanto em E_T quanto em η.

Além disso, é importante também analisar o desvio da estimação de energia obtida na etapa rápida do *trigger* para seu valor real, na Figura 23 é possível encontrar o *box-plot* de α, definido na Eq. (4.1), nos espaços de fase E_T e η quando consideramos a energia real do evento e a energia estimada estimada, sendo o caso ideal $\alpha = 1$.

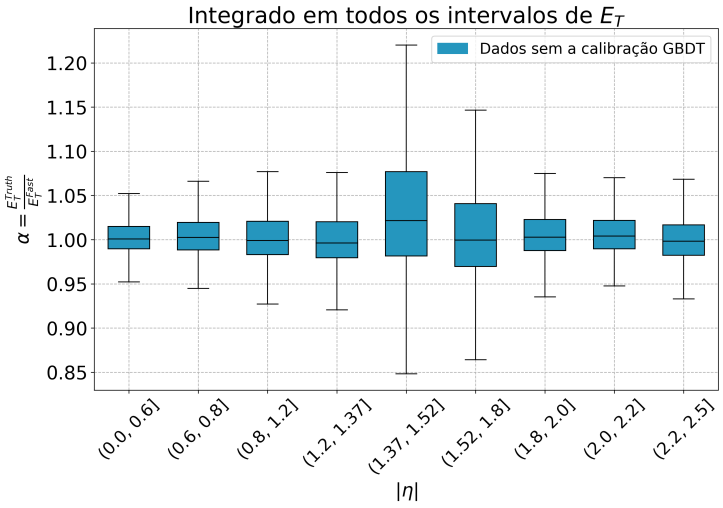
Observa-se na Figura 22(a) que no conjunto E_T, a dispersão é mais significativa em baixas energias, com uma notável diferença entre a mediana e o valor ideal. Enquanto isso, na

Figura 22(b), é possível verificar que para o parâmetro η, é evidente que há uma dispersão considerável no intervalo [1.37, 1.52), com a mediana substancialmente desviada de seu valor ideal, já para o intervalo (1.52, 1.8] há uma grande dispersão.

Figura 23 – Diagramas de caixa de α com o fenômeno de *pileup*, integrados em E_T e η para o conjunto de dados, antes da calibração.



(a) Diagrama de caixa para os intervalos de E_T. Pode-se verificar que a dispersão é maior para as baixas energias.



(b) Diagrama de caixa para os intervalos de | η |. É possível observar que há uma maior dispersão na região de 1.37 até 1.52 | η |

Fonte: Autoria Própria

Por fim, verificaremos o desempenho do modelo no ambiente computacional do ATLAS. Para isso, utilizou-se uma *Sample A* de validação composta por eventos de decaimento $Z \to e^+e^-$, processados por toda a cadeia de simulação e trigger. Em seguida, avaliou-se os histogramas de seleção dos elétrons e a eficiência desta seleção, comparando-os à referência previamente estabelecida para identificar possíveis desvios no desempenho do *trigger*.

4.3 Definição de parâmetros da GBDT

Para a implementação da GBDT foi o utilizado o *framework* LightGBM (KE et al., 2017). Há alguns trabalhos anteriores da colaboração ATLAS que já utilizam esse *framework* para a calibração de energia na reconstrução *offline*, isto é, após os eventos terem sido selecionados e salvos em mídia permanente para posterior análise física pelo *trigger*, bem como na etapa precisa do HLT (ATLAS Collaboration, 2019).

Como mencionado anteriormente, este trabalho visa investigar a utilização dos anéis de deposição de energia (*rings*) como entradas para a calibração GBDT. Para comparação, são empregados atributos tradicionalmente calculados do chuveiro de partículas (*shower variables*). As informações das variáveis de entrada para ambos os casos são apresentadas nas Tabelas 4 e 5.

Tabela 4 – Descrição das variáveis de entrada e saída alvo para a abordagem de anéis.

Variável	Descrição
$E_{T\ \text{cluster}}$	Energia medida do cluster
η_{cluster}	Pseudo-rapidez do centro do cluster
Anéis	Anéis concêntricos montados centrados na célula mais energética
Saída alvo	Descrição da saída alvo
α	$\frac{E_T^{\text{Truth}}}{E_T^{\text{Fast}}}$

Tabela 5 – Descrição das variáveis de entrada e saída alvo para a abordagem de chuveiros.

Variável	Descrição
$E_{T\ \text{cluster}}$	Energia medida do cluster
η_{cluster}	Pseudo-rapidez do centro do cluster
$\frac{E1_{\text{raw}}}{E2_{\text{raw}}}$	Razão entre as energias da EM1 e EM2
$E_{\text{raw}} = E1_{\text{raw}} + E2_{\text{raw}} + E3_{\text{raw}}$	Somatório de energia
$\frac{E0_{\text{raw}}}{E1_{\text{raw}}}$	Razão entre as energias da PS e EM1
$\frac{E_{\text{Tile1}}}{E_{\text{raw}}}$	Razão entre as energias da HAD1 e EM
Saída alvo	Descrição da saída alvo
α	$\frac{E_T^{\text{Truth}}}{E_T^{\text{Fast}}}$

Os eventos disponíveis foram divididos em conjuntos de treino, validação e teste com respectivamente 50%, 20% e 30% do total de exemplos, para cada faixa de $|\eta|$ e E_T. Além disso, os hiperparâmetros das GBDT, que foram escolhidos após testes exaustivos, são:

- Função de perda L1;

- 60 folhas em cada árvore;
- 2000 rodadas de *boosting*;
- Taxa de aprendizagem 0,05;
- Razão de *bagging* 0,8;
- Parada antecipada de 5 rodadas sem mudança no conjunto de validação;
- Validação cruzada com 5 partições;

Neste capítulo, foi detalhada a metodologia empregada para a construção do modelo GBDT e para a verificação de seu desempenho no ambiente do experimento ATLAS. O capítulo seguinte apresentará os resultados obtidos a partir da aplicação desta metodologia, bem como a análise correspondente.

5 Resultados

Este capítulo apresenta e discute os resultados obtidos ao longo da pesquisa, com base na metodologia descrita no capítulo anterior. Na seção 5.1, são apresentados os resultados obtidos a partir dos dados das simulações de Monte Carlo, incluindo gráficos, tabelas e figuras que ilustram as principais descobertas. Na seção 5.2, é feita uma análise do limiar de seleção e do desempenho do modelo nas cadeias de seleção do ATLAS. Por fim, na seção 5.3, será realizada uma interpretação mais profunda dos resultados, proporcionando uma compreensão mais ampla e contextualizada dos achados.

5.1 Resultados Obtidos no Desenvolvimento e Testes do Sistema Proposto

Na Figura 24, é possível visualizar o perfil da energia normalizada dos anéis. A distribuição de energia em cada anel foi representada utilizando um gráfico *boxplot*, para ilustrar a mediana, os quartis e a dispersão dos dados. A normalização aplicada seguiu o mesmo procedimento utilizado no pré-processamento para utilização do algoritmo NeuralRinger, onde os valores das energias de cada anel são normalizados para operar dentro da faixa dinâmica das redes neurais, como segue:

$$r_k = \frac{R_k}{\left|\sum_{i=1}^{100} R_i\right|}, \quad \forall k \in \{1, 2, \ldots, 100\}, \tag{5.1}$$

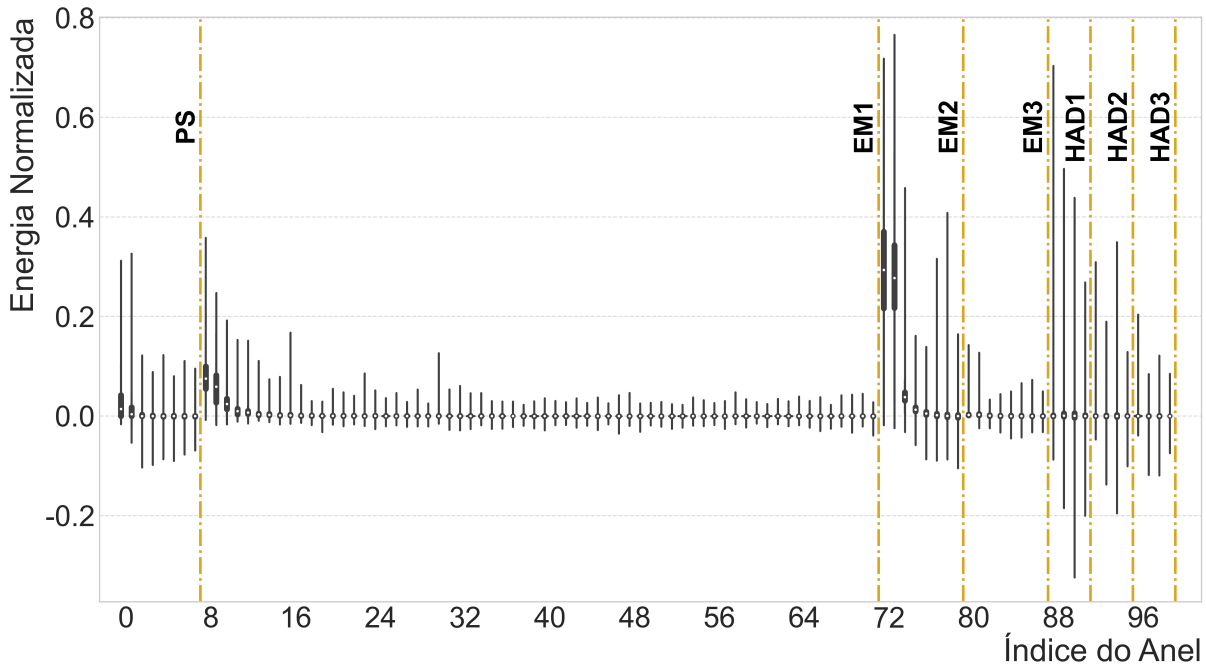
em que R_k representa o valor da energia do k-ésimo anel e r_k seu respectivo valor normalizado.

Observa-se o padrão característico de deposição de energia para elétrons. A maior parte da energia está concentrada nas camadas do calorímetro eletromagnético, com picos evidentes em EM1 e EM2. Nota-se que a camada EM1 apresenta pouca variação nos valores de energia depositada. Em contraste, a camada que mais demonstra variação é a EM2, o que sugere uma maior flutuação na contenção do chuveiro eletromagnético nesta região mais profunda. Como esperado para partículas eletromagnéticas, a deposição de energia nas camadas hadrônicas (HAD1, HAD2 e HAD3) é mínima, apesar da alta dispersão que pode ter sido ocasionada pelo empilhamento de sinais.

Em um segundo momento, analisaremos a dispersão e o deslocamento da mediana no espaço de fase η e E_T após o treinamento dos modelos com duas estratégias distintas para a seleção de parâmetros de entrada. Para isso, utilizaremos diagramas de caixa que são apresentados nas Figuras 25(a) e 25(b). Esses diagramas são integrados em ambos os espaços de fase, ou seja, a energia transversa E_T e a pseudo-rapidez η. Essas figuras permitem uma comparação visual das estimativas do parâmetro α em dois cenários distintos de calibração, bem como em um cenário sem calibração, oferecendo uma visão clara das melhorias e dos efeitos do processo de calibração.

Na Figura 25(b), observamos que a calibração que utiliza informações tanto do chuveiro quanto dos anéis resulta em uma significativa redução da dispersão das estimativas ao longo da maior parte da faixa de $|\eta|$. Essa redução na dispersão é um indicativo direto da eficácia

Figura 24 – Perfil médio da energia normalizada dos anéis



Fonte: Autoria Própria

da calibração em uniformizar as estimativas, tornando-as mais consistentes e menos sujeitas a variações indesejadas. A análise é consistente ao longo da faixa de pseudo-rapidez, mostrando que a calibração melhora a precisão das estimativas em diferentes regiões do espaço de fase.

Similarmente, a Figura 25(a) revela que a aplicação do método GBDT reduz a dispersão das estimativas em vários intervalos de energia transversa E_T, independentemente da estratégia de calibração empregada. Este resultado é relevante, pois demonstra que o GBDT tem um impacto positivo em diferentes condições de energia, contribuindo para a precisão das estimativas. Além da diminuição da dispersão, é notável que a mediana do parâmetro α tende a se aproximar de 1 após a calibração. Isso indica que, com a calibração, as estimativas de energia estão mais alinhadas com o valor verdadeiro, sugerindo uma melhora substancial na precisão das previsões.

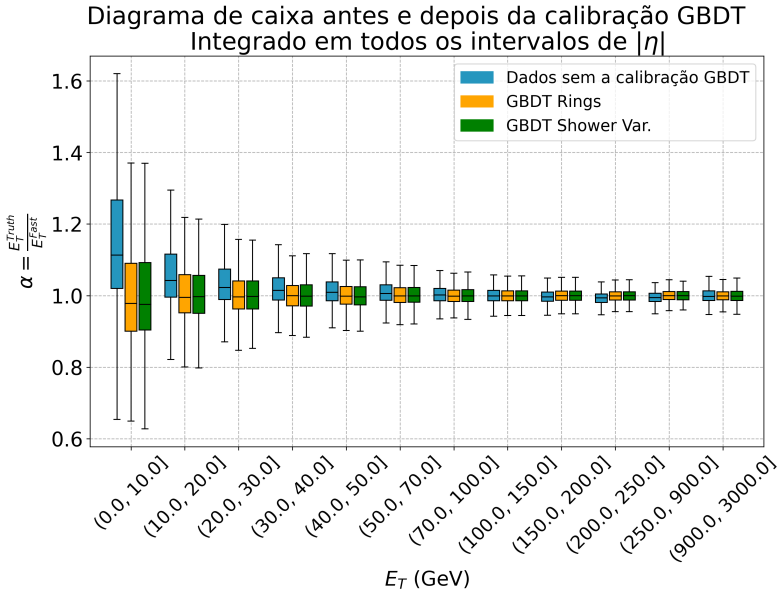
Para aprofundar a análise, as Figuras 26 e 27 fornecem uma visão detalhada das distâncias interquartílicas (IQR = Q3 - Q1)[1] para as faixas de energia e de pseudo-rapidez mencionadas anteriormente. Essas figuras ilustram como a dispersão das estimativas de energia é afetada pela calibração.

Além disso, é apresentada a razão normalizada da IQR em relação ao cenário sem calibração, facilitando a comparação entre os diferentes métodos de calibração e destacando a eficácia relativa de cada abordagem. Essas análises quantitativas são fundamentais para entender a extensão da melhoria proporcionada pela calibração e para avaliar a consistência das estimativas em diferentes

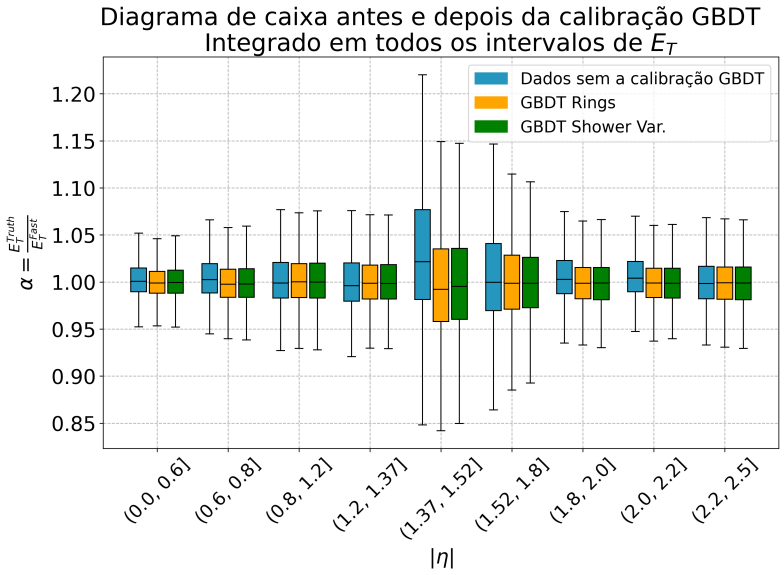
[1] O intervalo interquartil (IQR) é uma medida de dispersão que representa a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1). Os quartis dividem um conjunto de dados ordenado em quatro partes iguais: $Q1$ é o ponto abaixo do qual se encontram 25% dos dados, enquanto Q3 é o ponto abaixo do qual estão 75% dos dados. O IQR é calculado como IQR = Q3 - Q1, e reflete a amplitude do intervalo central que contém 50% dos dados. Em um diagrama de caixa, o IQR é representado pela largura da caixa, ilustrando a dispersão dos dados ao redor da mediana.

condições experimentais.

Figura 25 – Diagramas de caixa de α integrados em E_T e $|\eta|$ para o conjunto de dados, antes e após a calibração.



(a) Diagrama de caixa para os intervalos de E_T.



(b) Diagrama de caixa para os intervalos de $|\eta|$.

Fonte: Autoria Própria

Ao verificar a razão normalizada em relação ao caso sem calibração, observa-se que o uso de anéis como entrada resulta em um desempenho superior, com melhorias significativas em quase todas as faixas de segmentação, tanto em $|\eta|$ quanto em E_T. A única exceção foi na primeira faixa de segmentação em E_T, onde o desempenho foi inferior ao uso das variáveis de chuveiro. No entanto, ambas as abordagens apresentaram resultados melhores do que o caso sem calibração.

Ademais, na primeira faixa de E_T (0 a 10 GeV) e na faixa de (1,37 a 1,52 η), as duas onde houve a pior performance, foi possível obter melhorias de 25% e 20%, respectivamente.

Figura 26 – IQR de α para a segmentação em $|\eta|$.

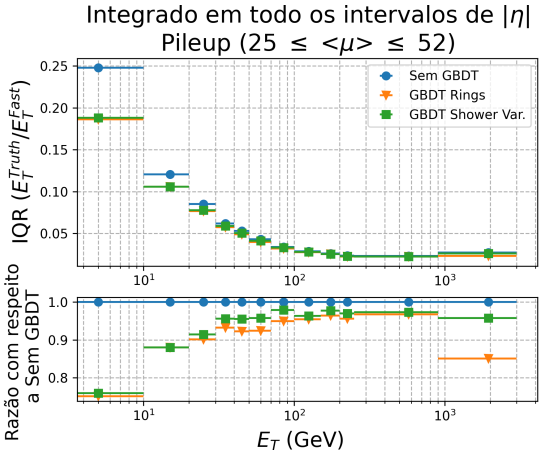
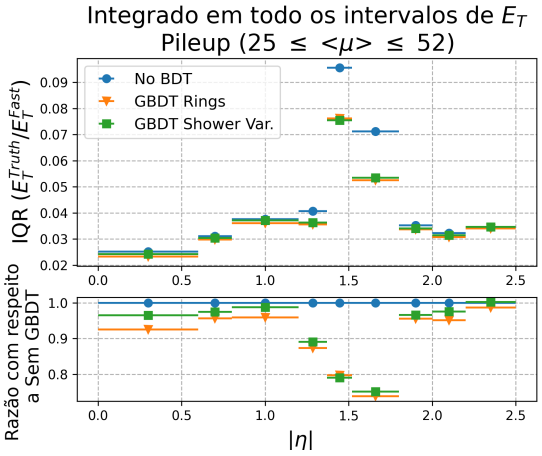


Figura 27 – IQR de α para a segmentação em E_T.



Fonte: Autoria Própria

Outra análise que pode ser feita para avaliar o impacto da calibração é calcular o erro relativo de estimação (Eq. (5.2)). Para isso, podemos considerar duas situações: uma em que utilizamos a energia estimada e outra em que utilizamos os valores de energia calibrada. O erro relativo de estimação, e_r, é uma medida que nos permite avaliar a precisão da calibração em relação aos valores verdadeiros. Para visualizar a distribuição dos erros relativos de estimação, podemos construir histogramas que mostrem a frequência dos valores de e_r para diferentes intervalos. Esses histogramas são úteis para identificar possíveis vieses na calibração e para verificar se os erros estão concentrados em torno de zero, o que indicaria uma boa calibração.

$$e_r = \frac{E_T^{(\text{Fast})} - E_T^{(\text{Truth})}}{E_T^{(\text{truth})}}. \tag{5.2}$$

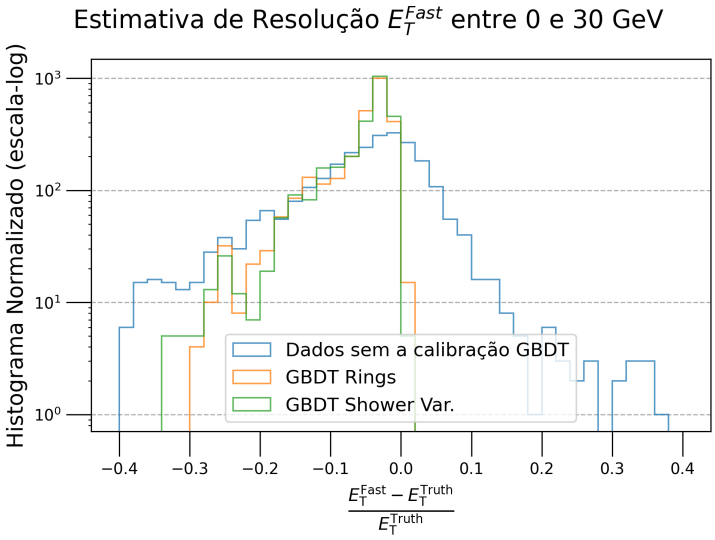
Na Figura 28 apresenta-se o histograma do erro relativo de estimação e_r, utilizando escala logarítmica no eixo y, para eventos com energias na faixa de 0 a 30 GeV. Notavelmente, observamos que a aplicação da calibração, seja utilizando *Shower Variables* ou *Rings*, resulta em um aumento significativo na quantidade de eventos com erro próximo a zero. Além disso, a dispersão dos valores de e_r também é visivelmente reduzida. Esses resultados indicam que a calibração dos parâmetros, independentemente da abordagem utilizada, tem um impacto positivo na precisão das estimativas de energia.

Na Figura 29, são apresentados dois gráficos de dispersão, onde o eixo horizontal representa o valor real da energia, enquanto o eixo vertical mostra o valor estimado, tanto antes quanto após a calibração. Especificamente, a calibração utilizada foi baseada nos anéis como estratégia de entrada para o modelo.

Nos gráficos de dispersão, observa-se claramente que, para energias mais baixas, as estimativas não calibradas apresentam erros significativos, resultando em uma dispersão consideravelmente maior dos pontos. Isso indica uma inconsistência nas previsões iniciais, especialmente em faixas de energia mais baixa, onde os valores estimados tendem a divergir substancialmente dos valores reais.

Após a implementação da calibração, notamos uma correção substancial desses erros. A

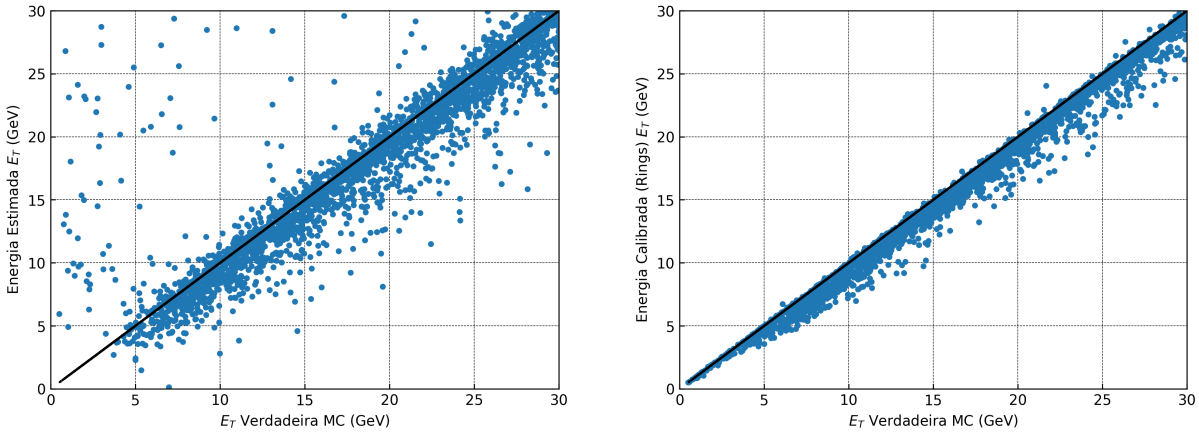
Figura 28 – Histograma de resolução da estimação para E_T entre 0 e 30 GeV.



Fonte: Autoria Própria

dispersão dos dados em torno da linha de identidade (onde o valor estimado é igual ao valor real) é visivelmente reduzida. Essa redução na dispersão é um indicativo direto da melhoria na precisão das estimativas proporcionada pela calibração.

Figura 29 – Gráfico de dispersão mostrando a comparação entre os valores reais e estimados de energia, sem calibração(a) e com calibração utilizando como estratégia de entrada baseada nos anéis (b) .



(a) **(b)**

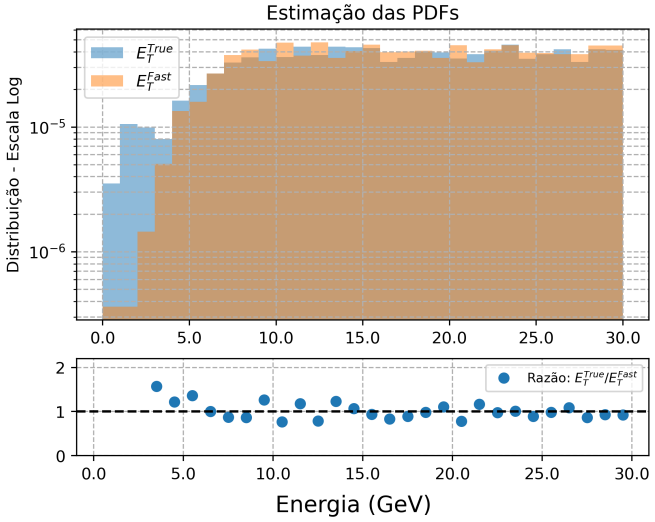
Fonte: Autoria Própria

Além disso, para essa faixa de energia foi possível obter uma melhora no MAPE de 20,6% para 5,8%.

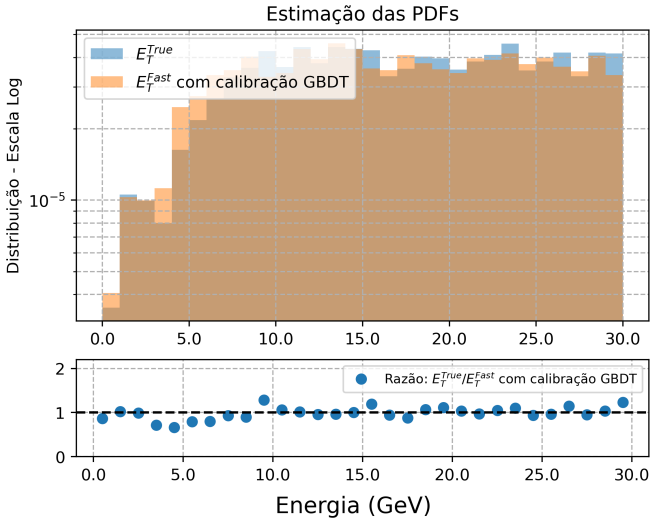
Ao examinar detalhadamente os resultados para a faixa de energia entre 0 e 30 GeV, nota-se na Figura 30 que o histograma da energia, antes da calibração com o uso do GBDT (Figura 30(a)), apresenta um descasamento significativo em relação ao histograma dos valores verdadeiros de energia. Este desalinhamento indica uma discrepância inicial na estimativa de energia. No entanto,

após a aplicação do processo de calibração, observamos na Figura 30(b) uma redução substancial dessa discrepância, evidenciando a eficácia do método GBDT em ajustar as estimativas de energia para valores mais próximos dos reais. Esse comportamento de melhoria consistente é também percebido em outras faixas de energia, destacando a robustez do método ao longo de um amplo espectro de energias analisadas.

Figura 30 – Comparação entre os histogramas da energia verdadeira versus a energia estimada na etapa rápida (a) antes e (b) depois da calibração com GBDT.



(a)



(b)

Fonte: Autoria Própria

5.2 Avaliação do Modelo na Cadeia de Seleção de Eventos no ATLAS

5.2.1 Avaliação dos Limiares de Seleção

Após o desenvolvimento do modelo e a realização dos testes com os conjuntos de simulação, é essencial avaliá-lo no ambiente real de seleção do ATLAS, com o objetivo de verificar sua eficácia em condições operacionais e conferir se os ajustes realizados na fase de simulação se refletem no desempenho durante a seleção de eventos.

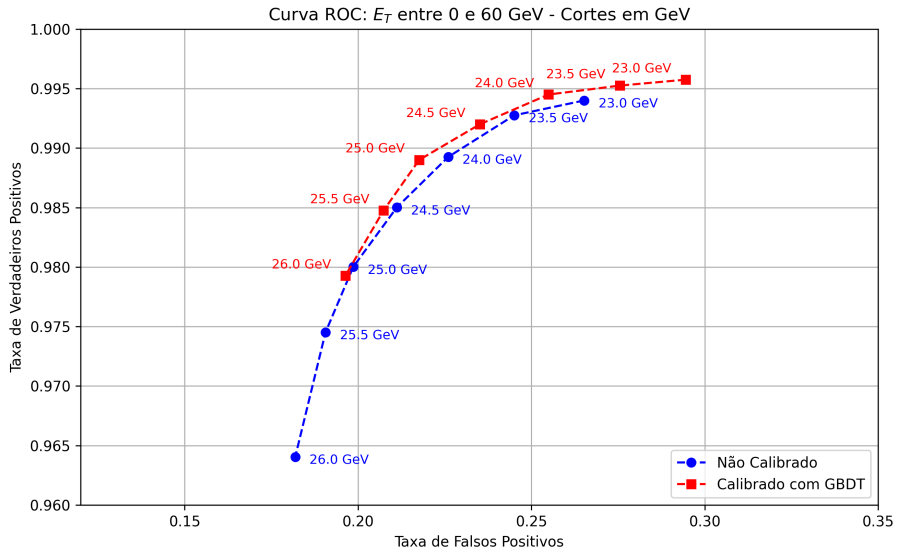
Como citado anteriormente, a ativação das cadeias de *trigger* utiliza um patamar de energia mínima. Por exemplo, as cadeias E26 e E60 selecionam candidatos a elétrons com energias superiores a 26 GeV e 60 GeV, respectivamente. Antes de avaliar o modelo, entretanto, é fundamental verificar se os limiares energéticos estão calibrados corretamente, garantindo que os critérios de seleção sejam aplicados conforme projetado. Essa verificação será realizada com o conjunto de testes, assegurando a consistência entre os parâmetros definidos e o comportamento observado no sistema.

Para acomodar os erros de estimação de energia que existem atualmente na etapa rápida, o patamar de seleção das cadeias (λ_{Fast}) é definido 3 GeV abaixo do valor ideal para o menu (λ_{Menu}), sendo definido como:

$$\lambda_{\text{Fast}} = \lambda_{\text{Menu}} - 3. \tag{5.3}$$

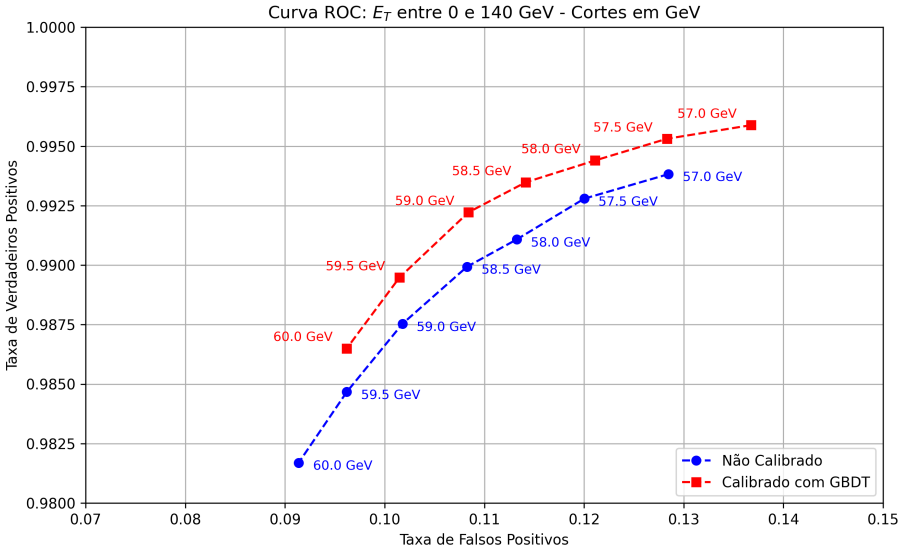
Para analisar os efeitos da calibração na seleção de eventos, é fundamental considerar diferentes limiares de seleção e avaliar como esses limiares influenciam os resultados. Nas Figuras 31 e 32, são apresentadas curvas ROC para ambos os cenários (E26 e E60), permitindo visualizar o impacto da calibração no desempenho do modelo.

Figura 31 – Curva ROC referente à seleção de candidatos a elétrons com energia superior a 26 GeV. A análise considera eventos com energia total de até 60 GeV.



Fonte: Autoria Própria

Figura 32 – Curva ROC referente à seleção de candidatos a elétrons com energia superior a 60 GeV. A análise considera eventos com energia total de até 140 GeV.



Fonte: Autoria Própria

Percebe-se que é possível diminuir a taxa de falsos positivos de modo a manter a mesma probabilidade de detecção original, reduzindo consideravelmente a aceitação de elétrons fora da faixa desejada. Nesse caso, fazendo:

$$\lambda_{\text{Fast}} = \lambda_{\text{Menu}} - 2. \tag{5.4}$$

A Tabela 6 apresenta os valores de elétrons fora da faixa desejada para diferentes cadeias de *trigger*. Observa-se que as reduções nas taxas de falso alarme (PF) foram modestas, atingindo 3,9% para E26 (de 26,52% para 25,49%) e 5,8% para E60 (de 12,85% para 12,11%).

Tabela 6 – Probabilidade de Detecção (PD) e Falso Alarme (PF) para diferentes cadeias de trigger.

Cadeia do *Trigger*	Sem GBDT		GBDT - Anéis	
	PF (%)	PD (%)	PF (%)	PD (%)
E26	26.52	96.37	25.49	96.77
E60	12.85	95.87	12.11	96.06

Fonte: Autoria Própria

5.2.2 Análise nas Cadeias de Seleção

5.2.2.1 Arquitetura e Integração da Calibração

Para analisar o desempenho do modelo nas cadeias de seleção do *trigger* de elétrons do ATLAS, foi preciso integrá-lo ao *framework* Athena. A lógica de processamento deste *framework* é predominantemente escrita em C++, visando à máxima eficiência computacional, enquanto a configuração de alto nível e a definição de parâmetros são gerenciadas por *scripts* Python. Nesse contexto, a nova calibração foi inserida como um componente no algoritmo `T2CaloEgammaReFastAlgo`, ou `ReFastAlgo`, que opera no segundo nível do *trigger* (HLT) e reconstrói os *clusters* eletromagnéticos a partir das Regiões de Interesse (*RoIs*).

A fim de garantir uma separação clara de responsabilidades e facilitar futuras atualizações, a lógica da calibração não foi codificada diretamente no `ReFastAlgo`. Em vez disso, essa lógica foi encapsulada em uma ferramenta dedicada do tipo `AsgTool` [2], a `TrigFastCalibWithRings`. Essa abordagem orientada a componentes é uma prática padrão no Athena e oferece vantagens substanciais: o `ReFastAlgo` delega a tarefa de calibração, tornando-se um "consumidor"do serviço, sem necessitar conhecer os detalhes internos do modelo de GBDT.

A ferramenta `TrigFastCalibWithRings` é inicializada carregando seus recursos de um arquivo ROOT (*framework* de análise e formato de arquivo de dados desenvolvido pelo CERN e amplamente utilizado na física de partículas (BRUN; RADEMAKERS, 1997)) externo, cujo caminho é um parâmetro configurável. Durante esta fase, o método `setupBDTFastCalo` lê os modelos de GBDT treinados, um histograma `TH2Poly` que particiona o espaço de fase em *bins* de pseudorapidez (η) e energia transversa (E_T), bem como as listas de variáveis de entrada para cada modelo. Essa abordagem permite que os modelos de calibração sejam atualizados e substituídos de forma independente, sem a necessidade de recompilar o código-fonte do *trigger*, o que agiliza significativamente o ciclo de pesquisa e desenvolvimento.

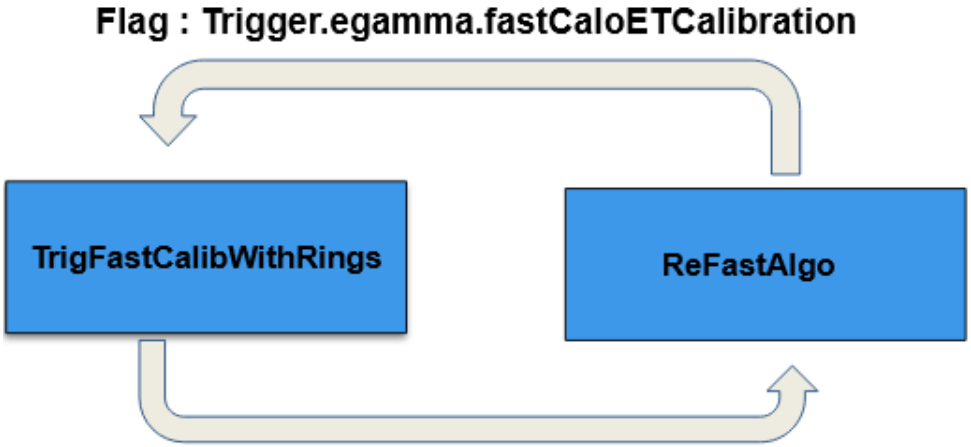
A ativação da calibração é controlada por uma propriedade configurável, a *flag* que têm nome `Trigger.egamma.fastCaloETCalibration`. Este mecanismo de controle é uma característica fundamental do *framework* Athena, que permite uma ponte entre o código C++ compilado e os *scripts* de configuração em Python.

Quando ativada, o `ReFastAlgo` instancia e invoca a ferramenta `TrigFastCalibWithRings` durante o processamento de cada evento. Esta, por sua vez, executa o método `makeCalibWRings`, que aplica o modelo de GBDT para obter um fator de correção e, consequentemente, ajustar o valor da energia transversa (E_T) do *cluster*. O fluxo de execução, condicionado pela *flag*, está ilustrado no diagrama da Figura 33.

É possível encontrar a implementação completa no repositório do ATLAS (*ATLAS Code Browser*).

[2] A `AsgTool` (*Analysis Software Group Tool*) é uma classe de componente fundamental no *framework* de *software* do experimento ATLAS. Seu *design* de "uso duplo"(*dual-use*) permite que a mesma ferramenta opere de forma idêntica tanto no ambiente de *software* completo (Athena) quanto em análises mais leves baseadas em *ROOT*, garantindo consistência, reutilização de código e manutenibilidade em todo o fluxo de trabalho da análise de física (BASAGLIA et al., 2015).

Figura 33 – Diagrama de Implementação da Calibração no Athena

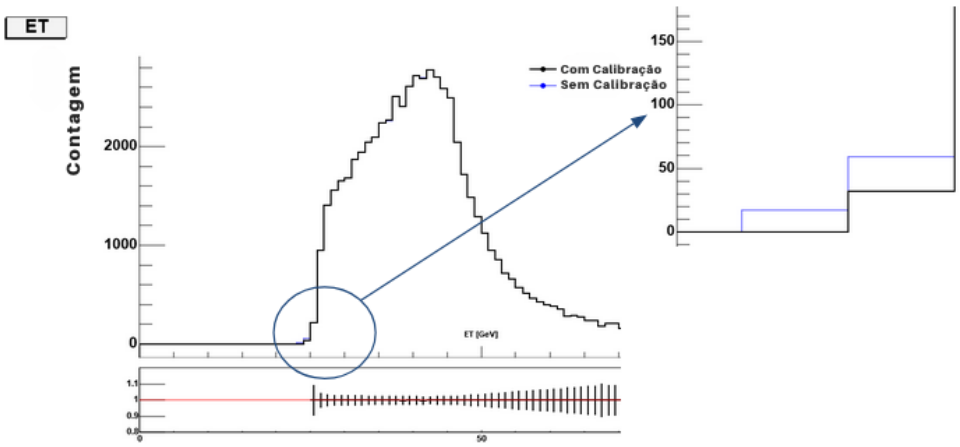


Fonte: Autoria Própria

5.2.2.2 Resultados

Como citado anteriormente os testes nas cadeias de seleção foram conduzidos utilizando o *Sample-A* (decaimentos $Z \rightarrow e^{+}e^{-}$). Além disso, foi assumida uma diferença de 1.5 entre os parâmetros λ_{Fast} e λ_{Menu}, com base em estudos anteriores (ALVES et al., 2023). Para avaliar o impacto da calibração nas cadeias de seleção do ATLAS, foram analisadas as distribuições de energia e a eficiência de detecção dos candidatos a elétrons nas cadeias E26 e E60. As Figuras 34 e 35 mostram os histogramas de energia antes e após a calibração, enquanto as Figuras 36 e 37 exibem as curvas de eficiência correspondentes.

Figura 34 – Distribuição de energia para a cadeia E26: dados não calibrados (azul) e calibrados (preto).

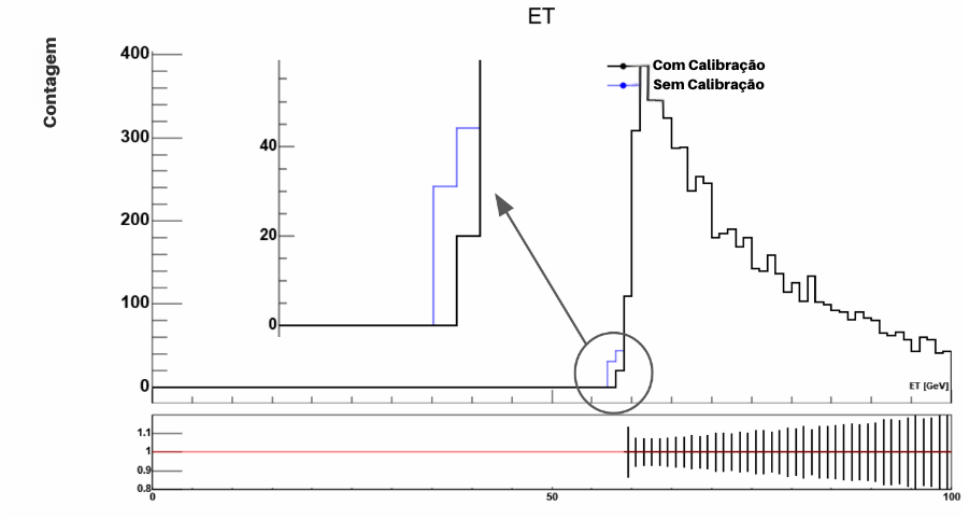


Fonte: Colaboração ATLAS (Adaptado)

Nas distribuições de energia (Figuras 34 e 35), a calibração reduz a quantidade de eventos próximos aos limiares de 26 GeV e 60 GeV, indicando uma filtragem mais seletiva de sinais de baixa energia.

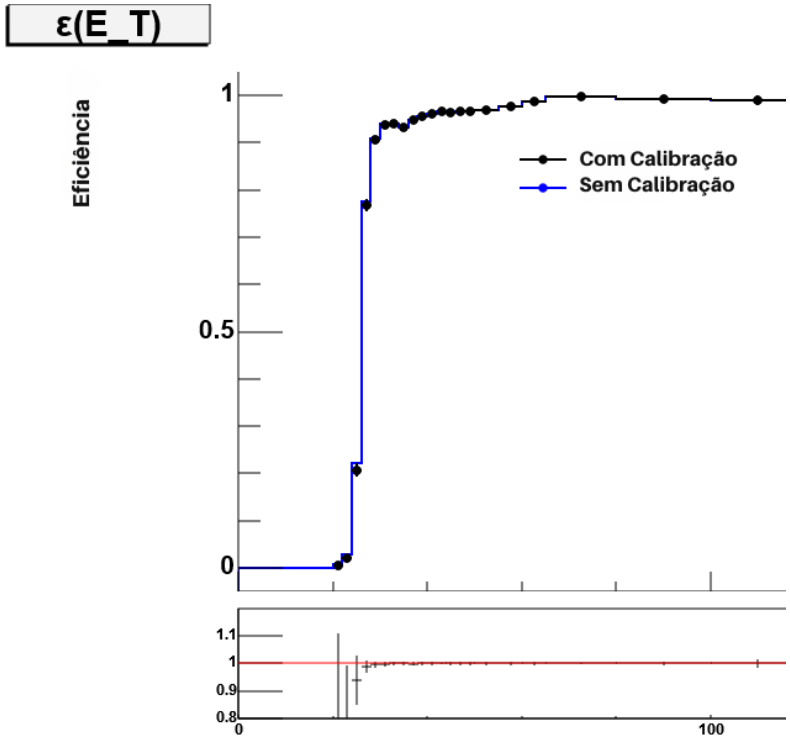
Essa redução não compromete a eficiência de detecção para ambas as cadeias, conforme

Figura 35 – Distribuição de energia para a cadeia E60: dados não calibrados (azul) e calibrados (preto).



Fonte: Colaboração ATLAS (Adaptado)

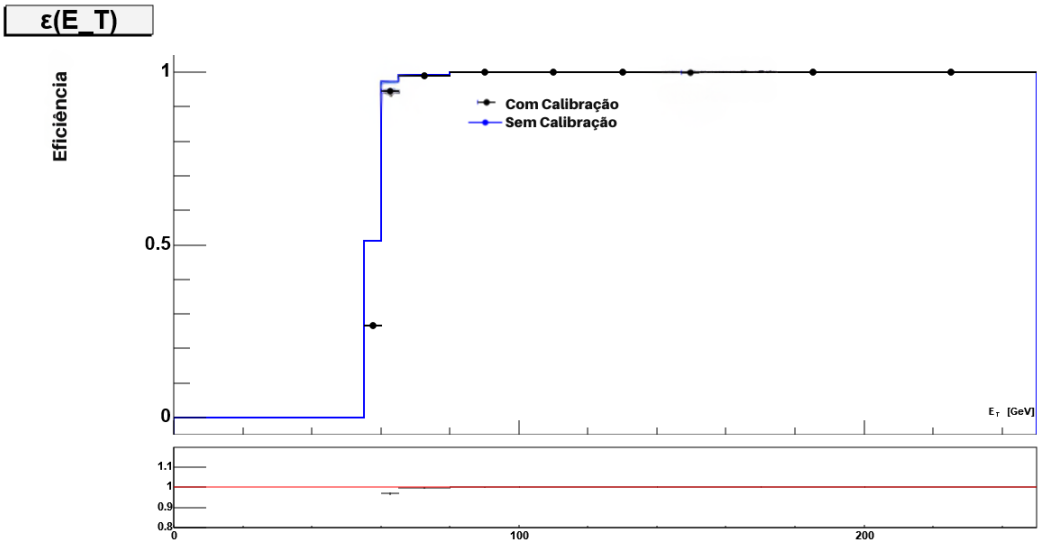
Figura 36 – Eficiência de detecção para a cadeia E26.



Fonte: Colaboração ATLAS (Adaptado)

mostrado nas Figuras 36 e 37. Na Figura 37, a queda de eficiência observada entre 55 e 60 GeV corresponde a eventos abaixo do limiar nominal da cadeia E60 ($\lambda_{\text{Menu}} = 60$ GeV), que não são alvo da seleção. Como a cadeia E60 é projetada para aceitar eventos acima de 60 GeV, a redução nessa faixa reflete a supressão de sinais próximos ao limiar rápido ($\lambda_{\text{Fast}} = 57$ GeV), ajustando-se à estratégia de seleção. Acima de 60 GeV, a eficiência mantém-se estável, garantindo a detecção confiável dos eventos de interesse. A supressão de eventos próximos aos limiares sugere uma diminuição na aceitação de ruído de fundo, enquanto a estabilidade nas curvas

Figura 37 – Eficiência de detecção para a cadeia E60.



Fonte: Colaboração ATLAS (Adaptado)

de eficiência confirma a preservação da capacidade de identificar elétrons acima de 60 GeV. Esse equilíbrio é crítico para otimizar as taxas de aquisição do sistema de *trigger*, garantindo a eficiência operacional do experimento.

5.3 Análise dos Resultados Obtidos

Os resultados apresentados nesta seção referem-se à avaliação do modelo no conjunto de teste, composto por dados de simulação não utilizados durante o treinamento. As medidas quantitativas indicam uma melhoria na precisão das estimativas de energia após a calibração. No conjunto de teste, observou-se uma redução no erro médio absoluto percentual (MAPE) de 20,6% para 5,8% na faixa de 0 a 30 GeV. A dispersão das estimativas também foi reduzida, conforme evidenciado pelos diagramas de caixa nas Figuras 25b e 25a, que mostram uma distribuição mais compacta dos resíduos após a calibração.

A análise dos diagramas de dispersão (Figura 29) destaca o efeito da calibração. Antes do ajuste, as estimativas para energias abaixo de 30 GeV apresentavam dispersão elevada, com desvios em relação à linha de identidade. Após a calibração, os dados se concentram mais próximos da relação ideal, principalmente em baixas energias.

5.3.1 Testes de Limiares e Curvas ROC

As curvas ROC (Figuras 31 e 32) revelaram que a calibração permite operar com margens mais estreitas entre os limiares rápido (λ_{Fast}) e nominal (λ_{Menu}). Para a cadeia E26, a taxa de falsos positivos (PF) foi reduzida de 26,52% para 25,49%, mantendo a probabilidade de detecção (PD) em 96,7%. Na E60, a PF diminuiu de 12,85% para 12,11%, com PD estável em 96,1%, como apresentado na Tabela 6.

5.3.2 Desempenho nas Cadeias de Seleção

A validação do modelo nas cadeias de *trigger* E26 e E60 mostrou redução na quantidade de eventos com energias próximas aos limiares inferiores (abaixo de 26 GeV e 60 GeV, respectivamente), conforme ilustrado nas Figuras 34 e 35. As curvas de eficiência (Figuras 36 e 37) indicam que a capacidade de identificação de elétrons genuínos permaneceu estável, com valores acima de 95% em ambas as cadeias.

Em síntese, a calibração demonstrou eficácia tanto em testes controlados quanto em condições operacionais, alinhando-se aos requisitos de precisão e eficiência do experimento ATLAS.

6 Conclusões

6.1 Conclusões

As pesquisas na área de física de partículas desenvolvidas no CERN visam compreender a natureza constitutiva da matéria e validar modelos teóricos. Para isso, foi construído o LHC, o maior acelerador de partículas em operação.

Nesse contexto, o experimento ATLAS, um dos detectores do LHC, utiliza um sistema de seleção online (*trigger*) para filtrar eventos relevantes em tempo real. Com o aumento da luminosidade do LHC, esse sistema necessita de atualizações para lidar com taxas de dados elevadas e com o empilhamento de sinais nos sensores.

O *trigger* de elétrons é importante, pois esses léptons podem indicar processos raros, como o bóson de Higgs, ou até processos que extrapolam as previsões do Modelo Padrão. Uma identificação correta requer a determinação precisa da energia dos elétrons.

Neste estudo, foi proposto um sistema de calibração de energia desenvolvido para a etapa rápida do HLT no *trigger* de elétrons. Para isso, empregou-se um modelo de aprendizado de máquina *ensemble* de árvores de decisão com reforço por gradiente (GBDT) para produzir um fator de calibração α para valores estimados de energia transversa.

Nos testes realizados com os dados obtidos por meio de simulações Monte Carlo, foi possível reduzir a dispersão dos dados de seu valor correto nos espaços de fase E_T e η, observando melhorias no IQR de até 25%. Além disso, para a faixa de baixas energias (0 a 30 GeV) houve uma melhora no MAPE de 20,6%.

Já nos testes com dados de validação, foi possível diminuir o limiar de seleção e obter uma redução no número de falsos positivos nas cadeias do *trigger*, mantendo a eficiência na seleção, reduzindo a demanda computacional e aumentando a eficiência global de seleção.

6.2 Trabalhos Futuros

Com a implementação do modelo e do sistema no software Athena do ATLAS, trabalhos futuros deverão concentrar-se na aplicação de métodos de otimização do GBDT, a fim de obter melhores resultados. Além disso, será fundamental ajustar os limiares de decisão nas cadeias do *trigger* para aprimorar a eficiência da seleção e reduzir ainda mais os falsos positivos. Adicionalmente, investigações futuras poderão explorar a integração de novas técnicas de aprendizado de máquina e a adaptação do sistema a diferentes condições operacionais, contribuindo para a melhoria contínua do desempenho global do experimento.

Referências

ALVES, A. et al. Calibração de energia para a seleção online de elétrons com alta taxa e utilizando um conjunto de Árvores de decisão com reforço por gradiente e informação especialista de calorimetria. In: *Anais do XVI Congresso Brasileiro de Inteligência Computacional*. SBIC, 2023. (CBIC 2023). Disponível em: <http://dx.doi.org/10.21528/CBIC2023-145>. 68

ARBUZOV, A. B. *Quantum Field Theory and the Electroweak Standard Model*. 2018. 24

ATLAS Collaboration. *ATLAS tile calorimeter: Technical Design Report*. CERN, 1997. Disponível em: <http://cds.cern.ch/record/331062>. 30

ATLAS Collaboration. The ATLAS Experiment at the CERN Large Hadron Collider. *JINST*, v. 3, p. S08003, 2008. Also published by CERN Geneva in 2010. Disponível em: <https://cds.cern.ch/record/1129811>. 26

ATLAS Collaboration. *Expected electron performance in the ATLAS experiment*. Geneva, 2011. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2011-006. Disponível em: <https://cds.cern.ch/record/1345327>. 32

ATLAS Collaboration. Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc. *Physics Letters B*, Elsevier BV, v. 716, n. 1, p. 1–29, set. 2012. ISSN 0370-2693. Disponível em: <http://dx.doi.org/10.1016/j.physletb.2012.08.020>. 24, 26

ATLAS Collaboration. *ATLAS Liquid Argon Calorimeter Phase-II Upgrade: Technical Design Report*. CERN Document Server, 2017. Disponível em: <http://cds.cern.ch/record/2285582>. 29

ATLAS Collaboration. Performance of the atlas trigger system in 2015. *The European Physical Journal C*, Springer Science and Business Media LLC, v. 77, n. 5, maio 2017. ISSN 1434-6052. Disponível em: <http://dx.doi.org/10.1140/epjc/s10052-017-4852-3>. 22

ATLAS Collaboration. *Technical Design Report for the Phase-II Upgrade of the ATLAS TDAQ System*. CERN Document Server, 2017. Disponível em: <http://cds.cern.ch/record/2285584>. 27

ATLAS Collaboration. Electron and photon energy calibration with the atlas detector using 2015–2016 lhc proton-proton collision data. *Journal of Instrumentation*, IOP Publishing, v. 14, n. 03, p. P03017–P03017, mar. 2019. ISSN 1748-0221. Disponível em: <http://dx.doi.org/10.1088/1748-0221/14/03/P03017>. 33, 35, 36, 37, 42, 53, 57

ATLAS Collaboration. Performance of electron and photon triggers in atlas during lhc run 2. *The European Physical Journal C*, Springer Science and Business Media LLC, v. 80, n. 1, jan. 2020. ISSN 1434-6052. Disponível em: <http://dx.doi.org/10.1140/epjc/s10052-019-7500-2>. 27, 36

ATLAS Collaboration. Performance of the atlas level-1 topological trigger in run 2. *The European Physical Journal C*, Springer Science and Business Media LLC, v. 82, n. 1, jan. 2022. ISSN 1434-6052. Disponível em: <http://dx.doi.org/10.1140/epjc/s10052-021-09807-0>. 22, 35

ATLAS Collaboration. The atlas trigger system for lhc run 3 and trigger performance in 2022. arXiv, 2024. Disponível em: <https://arxiv.org/abs/2401.06630>. 22, 32, 33, 34, 35

BARRAND, G. et al. Gaudi — a software architecture and framework for building hep data processing applications. *Computer Physics Communications*, Elsevier BV, v. 140, n. 1–2, p. 45–55, out. 2001. ISSN 0010-4655. Disponível em: <http://dx.doi.org/10.1016/S0010-4655(01)00254-5>. 35

BASAGLIA, T. et al. Experimental quantification of geant4 physicslist recommendations: methods and results. *Journal of Physics: Conference Series*, IOP Publishing, v. 664, n. 7, p. 072037, dez. 2015. ISSN 1742-6596. Disponível em: <http://dx.doi.org/10.1088/1742-6596/664/7/072037>. 67

BENTÉJAC, C.; CSÖRGŐ, A.; MARTÍNEZ-MUÑOZ, G. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, Springer Science and Business Media LLC, v. 54, n. 3, p. 1937–1967, ago. 2020. ISSN 1573-7462. Disponível em: <http://dx.doi.org/10.1007/s10462-020-09896-5>. 50

BETTINI, A. *Introduction to Elementary Particle Physics*. Cambridge University Press, 2014. ISBN 9781107279483. Disponível em: <http://dx.doi.org/10.1017/CBO9781107279483>. 21

BISHOP, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN 0387310738. 39, 40, 47

BRUN, R.; RADEMAKERS, F. Root — an object oriented data analysis framework. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, Elsevier BV, v. 389, n. 1–2, p. 81–86, abr. 1997. ISSN 0168-9002. Disponível em: <http://dx.doi.org/10.1016/S0168-9002(97)00048-X>. 67

BRÜNING, O.; BURKHARDT, H.; MYERS, S. The large hadron collider. *Progress in Particle and Nuclear Physics*, Elsevier BV, v. 67, n. 3, p. 705–734, jul. 2012. ISSN 0146-6410. Disponível em: <http://dx.doi.org/10.1016/j.ppnp.2012.03.001>. 21

COLLABORATION, A. *ATLAS Liquid Argon Calorimeter Phase-II Upgrade: Technical Design Report*. Geneva, 2017. Disponível em: <https://cds.cern.ch/record/2285582>. 36

EVANS, L.; BRYANT, P. Lhc machine. *Journal of Instrumentation*, IOP Publishing, v. 3, n. 08, p. S08001–S08001, ago. 2008. ISSN 1748-0221. Disponível em: <http://dx.doi.org/10.1088/1748-0221/3/08/S08001>. 21, 25, 28, 29, 30, 31

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 29, n. 5, out. 2001. ISSN 0090-5364. Disponível em: <http://dx.doi.org/10.1214/aos/1013203451>. 46

GOODFELLOW, I. J.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. <http://www.deeplearningbook.org>. 39

GRAY, H. M. Future colliders for the high-energy frontier. *Reviews in Physics*, Elsevier BV, v. 6, p. 100053, jun. 2021. ISSN 2405-4283. Disponível em: <http://dx.doi.org/10.1016/j.revip.2021.100053>. 21

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning*. Springer New York, 2009. ISSN 2197-568X. ISBN 9780387848587. Disponível em: <http://dx.doi.org/10.1007/978-0-387-84858-7>. 43, 44, 45, 46, 47, 48, 49, 50

HAYKIN, S. O. *Neural Networks and Learning Machines*. 3. ed. Upper Saddle River, NJ: Pearson, 2008. 41

JAMES, G. et al. *An introduction to statistical learning*. 1. ed. Cham, Switzerland: Springer International Publishing, 2023. 41, 42, 43, 50

JONES, S. D. The atlas electron and photon trigger. *Journal of Physics: Conference Series*, IOP Publishing, v. 1162, p. 012037, jan. 2019. ISSN 1742-6596. Disponível em: <http://dx.doi.org/10.1088/1742-6596/1162/1/012037>. 34

JU, Y. et al. A model combining convolutional neural network and lightgbm algorithm for ultra-short-term wind power forecasting. *IEEE Access*, Institute of Electrical and Electronics Engineers (IEEE), v. 7, p. 28309–28318, 2019. ISSN 2169-3536. Disponível em: <http://dx.doi.org/10.1109/ACCESS.2019.2901920>. 50, 51

KAUR, R. et al. Melanoma classification using a novel deep convolutional neural network with dermoscopic images. *Sensors*, MDPI AG, v. 22, n. 3, p. 1134, fev. 2022. ISSN 1424-8220. Disponível em: <http://dx.doi.org/10.3390/s22031134>. 41

KE, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. In: GUYON, I. et al. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf>. 50, 51, 57

LUO, B. et al. Defect detection of metal sheets based on improved yolox algorithm. In: *2023 CAA Symposium on Fault Detection, Supervision and Safety for Technical Processes (SAFEPROCESS)*. IEEE, 2023. Disponível em: <http://dx.doi.org/10.1109/SAFEPROCESS58597.2023.10295669>. 41

MARIN, J. L. Energy reconstruction performance in the atlas tile calorimeter operating at high event rate conditions using lhc collision data. In: *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*. [S.l.: s.n.], 2020. p. 361–366. 22

MARZBAN, C. The roc curve and the area under it as performance measures. *Weather and Forecasting*, American Meteorological Society, v. 19, n. 6, p. 1106–1114, dez. 2004. ISSN 0882-8156. Disponível em: <http://dx.doi.org/10.1175/825.1>. 41

MORAIS, A. P. et al. Deep Learning Searches for Vector-Like Leptons at the LHC and Electron/Muon Colliders. *Eur. Phys. J. C*, v. 83, n. 3, p. 232, 2023. 26 pages, 11 figures, 10 tables, Published version. Disponível em: <https://cds.cern.ch/record/2853391>. 21

ORELLANA, G. E. Projected atlas electron and photon trigger performance in run 3. In: *Proceedings of The Eighth Annual Conference on Large Hadron Collider Physics — PoS(LHCP2020)*. Sissa Medialab, 2020. (LHCP2020). Disponível em: <http://dx.doi.org/10.22323/1.382.0244>. 21

PINTO, J. V. da F.; ATLAS Collaboration. An ensemble of neural networks for online filtering implemented in the atlas trigger system. *Journal of Physics: Conference Series*, IOP Publishing, v. 1162, p. 012039, jan. 2019. ISSN 1742-6596. Disponível em: <http://dx.doi.org/10.1088/1742-6596/1162/1/012039>. 21, 34, 35

RIPLEY, B. D. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996. ISBN 9780511812651. Disponível em: <http://dx.doi.org/10.1017/CBO9780511812651>. 46

SEIXAS, J. et al. Neural second-level trigger system based on calorimetry. *Computer Physics Communications*, Elsevier BV, v. 95, n. 2–3, p. 143–157, jun. 1996. ISSN 0010-4655. Disponível em: <http://dx.doi.org/10.1016/0010-4655(96)00012-4>. 35

SIMAS FILHO, E. F. et al. Uma proposta para calibração do sistema online de seleção de eventos no detector atlas utilizando gradient boosted decision trees. In: *Anais do 15. Congresso Brasileiro de Inteligência Computacional*. SBIC, 2021. (CBIC 2021). Disponível em: <http://dx.doi.org/10.21528/CBIC2021-149>. 38, 53

SOTO, P. C. Development of the ATLAS Liquid Argon Calorimeter Readout Electronics for the HL-LHC. 2025. Disponível em: <https://cds.cern.ch/record/2930866>. 21

SOYEZ, G. Pileup mitigation at the lhc: A theorist's view. *Physics Reports*, Elsevier BV, v. 803, p. 1–158, abr. 2019. ISSN 0370-1573. Disponível em: <http://dx.doi.org/10.1016/j.physrep.2019.01.007>. 54

WIGMANS, R. *Calorimetry: Energy Measurement in Particle Physics*. Oxford University PressOxford, 2017. ISBN 9780191828652. Disponível em: <http://dx.doi.org/10.1093/oso/9780198786351.001.0001>. 21, 22, 27, 28

A Trabalhos Publicados

A.1: Artigos Publicados em Anais de Congressos e Simpósios

1. ALVES, Arthur; SILVA, Paulo; SIMAS FILHO, Eduardo; FARIAS, Paulo César; MARIN, Juan; SEIXAS, José Manoel; SOUZA, Edmar; LAFORGE, Bertrand. Calibração de Energia para a Seleção Online de Elétrons com Alta Taxa e Utilizando um Conjunto de Árvores de Decisão com Reforço por Gradiente e Informação Especialista de Calorimetria. In: *Anais do XVI Congresso Brasileiro de Inteligência Computacional (CBIC 2023)*. São Paulo: SBIC, 2023. p. 145. Disponível em: <http://dx.doi.org/10.21528/CBIC2023-145>.

   - **Resumo**

   Em física experimental de altas energias, é preciso lidar com um grande volume de informações, sendo grande parte delas proveniente do ruído de fundo que dificulta a caracterização dos fenômenos de interesse particular de um dado experimento. Deste modo, é necessário um complexo processo de seleção online de eventos (trigger). No ATLAS, maior experimento do LHC (Large Hadron Collider), o sistema de trigger opera em duas etapas de seleção sequenciais, denominadas primeiro e alto nível. No caso de elétrons, importantes como mensageiros da nova física que se deseja observar, o sistema de trigger se apoia fortemente no sistema de calorimetria, que mede a energia da partícula incidente. Neste trabalho, é proposto um método de calibração de energia baseado em um conjunto de árvores de decisão com reforço por gradiente (*Gradient Boosted Decision Trees Ensemble* - GBDTE) para melhorar a acuidade da estimativa da energia na etapa rápida do trigger de alto nível do experimento ATLAS. Com esse método proposto, é possível reduzir os requisitos computacionais e aumentar a eficiência na seleção de partículas eletromagnéticas, como elétrons.

2. ALVES, Arthur; SILVA, Paulo; SIMAS FILHO, Eduardo; FARIAS, Paulo César; MARIN, Juan; SEIXAS, José Manoel; SOUZA, Edmar; LAFORGE, Bertrand. Calibração de Energia para Seleção Online de Eventos em Detector com Elevado Empilhamento de Sinais Utilizando um *Ensemble* de GBDTs. In: *Anais do XXV Congresso Brasileiro de Automática (CBA 2024)*. Rio de Janeiro: CBA, 2024.

   - **Resumo**

   Na física experimental de altas energias, lidar com um grande volume de dados é essencial para produzir informações relevantes, pois uma parcela significativa dos dados provém de ruído de fundo, dificultando a caracterização de fenômenos de interesse. Um processo complexo e sequencial de seleção de eventos on-line, conhecido como gatilho, é crucial para enfrentar esse desafio. No experimento ATLAS no Large Hadron Collider (LHC), o sistema de gatilho opera em dois estágios sequenciais: primeiro nível e alto nível. No caso dos elétrons, essenciais como mensageiros da nova física, o sistema de gatilho depende fortemente de calorímetros, que medem a energia das partículas incidentes. Este trabalho propõe um

método de calibração de energia baseado em árvores de decisão com gradiente reforçado (GBDT) para aumentar a precisão da estimativa de energia no gatilho de alto nível ATLAS. Este método pode reduzir os requisitos computacionais e aumentar a eficiência na seleção de partículas, incluindo elétrons, mesmo em cenários com presença de empilhamento.

A.2: Trabalhos Apresentados em Congressos e Simpósios

1. ALVES, Arthur; SILVA, Paulo; SIMAS FILHO, Eduardo; FARIAS, Paulo César; MARIN, Juan; SEIXAS, José Manoel; SOUZA, Edmar; LAFORGE, Bertrand. A Gradient-Boosted Decision Tree Ensemble Fed From Ring-Based Calorimeter Information For Trigger-Level Electron Energy Calibration Under Severe Pileup Conditions In The ATLAS Experiment. In: *IV Encontro de Primavera da SBF de 2024 (EPSBF 2024)*. Belo Horizonte: SBF, 2024.

   - **Resumo**

   Particle physics experiments handle a vast amount of data and require a sophisticated online event selection system (trigger) to retain the most relevant events. In the ATLAS detector at the Large Hadron Collider (LHC), this trigger system operates through two sequential steps: Level 1 and High-Level Triggers (HLT). The HLT in ATLAS is divided into Fast and Precision stages and utilizes information from calorimeters, muon detectors, and tracking systems. Executed on a parallel computer cluster, the HLT trigger menu relies on the estimated energy of particle candidates as a key parameter. In this study, we propose an energy calibration method using a gradient-boosted decision trees ensemble (GBDTE) for the FastCalo trigger step to enhance the accuracy of energy estimation for electron candidates. We apply a phase-space binning in pseudorapidity and transverse energy to better address the characteristics of calorimeter energy deposition profiles. The calibration utilizes ring-like calorimeter energy information as input variables, describing energy deposition profiles in terms of concentric rings built from original data, thus reducing information dimensionality while preserving most of the relevant details of particle shower development in the calorimeter system. Results from simulated single electron samples with high pileup levels demonstrate that the proposed approach is robust against pileup distortion. We present tests with a preliminary implementation in the ATLAS trigger and analysis software framework (Athena), focusing on energy estimation accuracy compared to Monte Carlo truth and offline reconstruction, the response time of the online implementation, and the effects on the full trigger chain. The proposed method can reduce computational requirements and increase efficiency in selecting electromagnetic particles.

2. ALVES, Arthur; SILVA, Paulo; SIMAS FILHO, Eduardo; FARIAS, Paulo César; MARIN, Juan; SEIXAS, José Manoel; SOUZA, Edmar; LAFORGE, Bertrand. Electrons Energy Calibration at the FastCalo Trigger Step using a Gradient Boosted Decision Tree Ensemble fed from Calorimeter Ring Sums Information. In: *7th ATLAS Machine Learning Workshop. CERN: ATLAS, 2024.*

   - **Resumo**

   In ATLAS, the electron trigger comprises hardware and software-based steps. Since 2017,

the High-Level Trigger (HLT) has been used for electron detection at the FastCalo step the NeuralRinger discriminator, a system based on pre-processing the calorimeter information into concentric rings and using such features to feed an ensemble of neural networks to distinguish between electrons and their usual hadronic backgrounds. The present work extends this approach by using the same energy profiles and a gradient-boosted decision trees ensemble (GBDTE) to improve the energy estimation accuracy in the FastCalo trigger step at the ATLAS HLT. The calorimeter information is formatted into concentric energy rings to produce a compacted and relevant set of variables for energy calibration. With the proposed method, it is possible to reduce computational requirements and increase efficiency in selecting electrons, performing a sharper cut in the energy of the identified object. For simulated single electron samples, it was possible to reduce the energy estimation error interquartile range (IQR) by approximately 10% in the barrel region (around $|\eta| = 1$) and in more than 20% in the low energy range (Et < 8 GeV).

3. ALVES, Arthur; SILVA, Paulo; SIMAS FILHO, Eduardo; FARIAS, Paulo César; MARIN, Juan; SEIXAS, José Manoel; SOUZA, Edmar; LAFORGE, Bertrand. Electrons Energy Calibration at the FastCalo Trigger Step using a Gradient Boosted Decision Tree Ensemble fed from Calorimeter Ring Sums Information. In: *8th ATLAS Machine Learning Workshop. CERN: ATLAS, 2025.*

   - **Resumo**

   In this work, we present the results from the deployment in Athena of an electron energy calibration method using gradient-boosted decision trees (GBDT) for the ATLAS FastCalo trigger step. The approach uses phase-space binning in pseudorapidity and transverse energy to better address different calorimeter shower profiles. Ring-like calorimeter features were estimated from the energy deposition profiles and used as input to the GBDT calibration. Such pre-processing reduces the information dimensionality and retains most of the relevant information on particle shower development in the calorimeter system. Tests on simulated single-electron samples under high pileup conditions demonstrated robustness to pileup effects. Implementation in the Athena framework was tested using Sample-A validation events ($Z \to e^+ e^-$ decays), and a reduction of fakes in the trigger chains was observed while preserving the true positive rate, lowering computational demands, and sharpening the efficiency turn-on. Should this technique prove effective, a future adaptation for the Global Trigger framework in the Run 4 L0 Trigger could be proposed.