CS460 - Machine Learning 2023

INSTRUCTOR: Dr. Subhankar Mishra
School of Computer Sciences, NISER

Project Mid-term presentation

Gradient Boosted Decision Trees and their application in improvising identification of decay of Higgs Bosons into pair of electrons.

Team members:
> Arpan Maity (Batch-19 SPS, NISER)
> Krishna Kant Parida (Batch-19 SPS, NISER)

What we had proposed in our project ?

Identification of Higgs Bosons produced either via gluon fusion or vector boson fusion channels decaying into electrons have a limiting branching fraction of 3.0×10^{-4} at 95% C.F. We are going to work to improvise the identification using GBDT algorithm by training a prediction model on the simulated and reconstructed dataset, and to work out event selection and categorization constraints related to the prepared prediction model.

Higgs boson coupling (interaction) with Fermions

Lagrangian of scalar Higgs field with Fermions:

$$\mathcal{L}_f = \frac{g_f v}{\sqrt{2}}\left(\bar{f}_L f_R + \bar{f}_R f_L\right) + \frac{g_f h}{\sqrt{2}}\left(\bar{f}_L f_R + \bar{f}_R f_L\right)$$

Vacuum Higgs field producing fermions

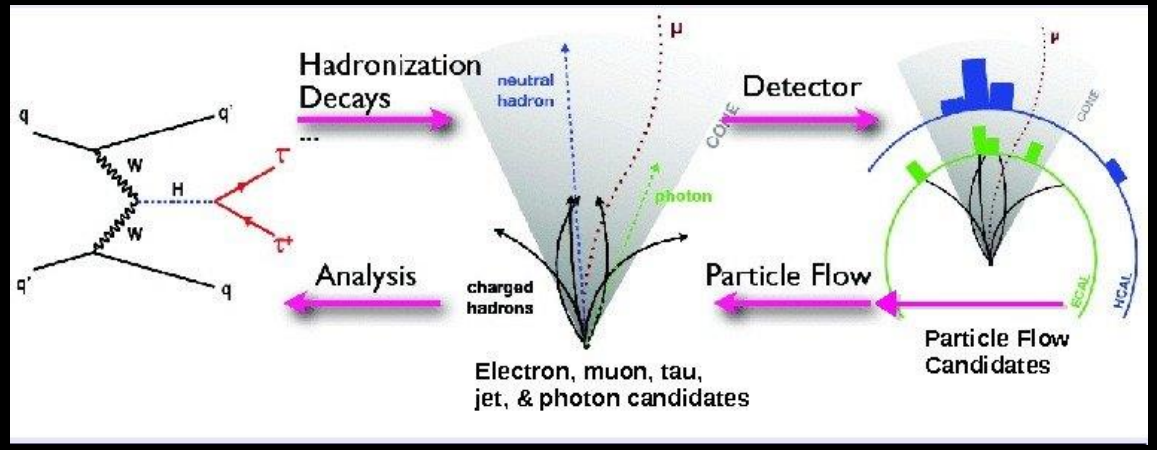
Coupling of fermions with Higgs Boson

$$m_f = \frac{g_f v}{\sqrt{2}}$$

Higgs coupling factor is directly proportional to the mass of the produced fermion

Higgs -> τ leptons [third generation fermions] -> 2012
-> muons [second generation fermions] -> 2014(CMS), 2019(ATLAS)
-> electrons [first generation fermions] -> Yet to be FOUND!

- CMS detector and particle flow algorithm were used for global event reconstruction of individual particles.



- Dielectron trigger signals were classified using various parametric thresholds for analysis strategy.

- Simulated sample dataset along with background were generated using Monte Carlo (MC) Techniques for Higgs production.

- Loose event preselection and VBF preselection procedure were applied for ensuring consistency with the required decay process detection.

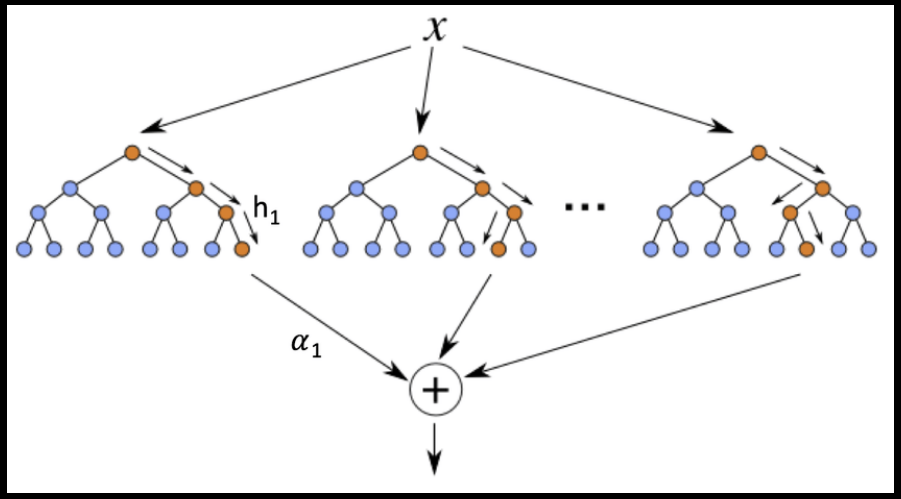
PYTHIA 8.3

- Event categorization primarily uses a boosted decision tree to discriminate the VBF based Higgs production as signal from background events.
- Categories are defined using the output scores of the classifying BDT, which are non-overlapping by construction.
- Each category to be analyzed is defined as per selections on a multivariate discriminant.
- A final model of an MVA-based classifier was trained to distinguish the signal events from the dominant Drell-Yan (DY) background signals.
- In each category, dielectron invariant mass distribution was fitted from which an attempt to extract an upper limit on the branching fraction was made.

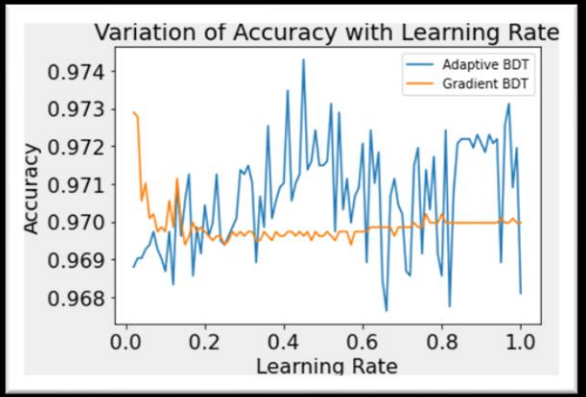
What we have done?

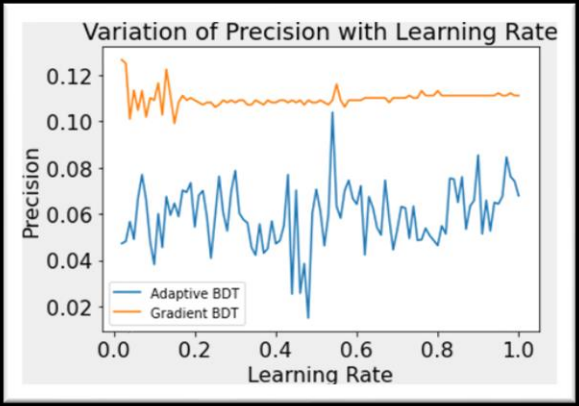
- The baseline is built upon classification of Higgs decay into a pair of electrons using a Gradient Boosting Decision Tree (GBDT).
- The GBDT is trained using a dataset simulated by Monte Carlo techniques using PYTHIA8. The root files generated by the simulator consist of both signal and background datasets.
- Specific feature branches from the trees representing VBF produced Higgs are selected for machine learning-based analysis.
- Data belonging to specific features, mainly representing the transverse momentum (p_T), angle (ϕ), and pseudorapidity (η) of each of the electrons, are taken into account and collected into an analyzable format like .csv file.

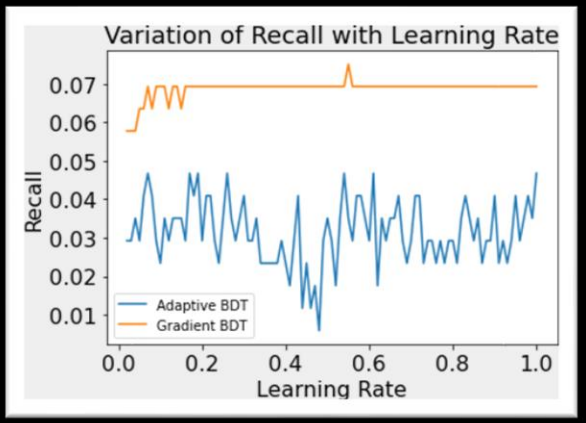
Schematic of a Boosted Decision Tree (BDT)

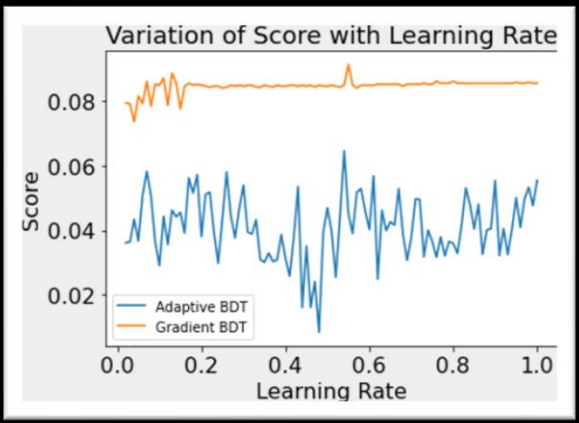


- The GBDT (from ScikitLearn) is trained with around 70% of the overall dataset, where 15% of the dataset is used for validation, and the final 15% is used for testing purposes. The section of data to be tested is chosen randomly from the overall dataset.
- Redundant features containing no significant data are removed, and conditions for event categorization are applied.
- The model is trained repeatedly over variation of learning rate and is tested using the test data from the dataset. The measures of accuracy, precision, recall, and BDT score are obtained from each testing procedure at different learning rates.
- Finally, the baseline analysis is done by comparison with adaptive boosting-based BDT (adaBoost classifier) in terms of their accuracy and BDT score. This helps in setting the learning rate of the GBDT and in search of possible tradeoffs associated with the learning rate.

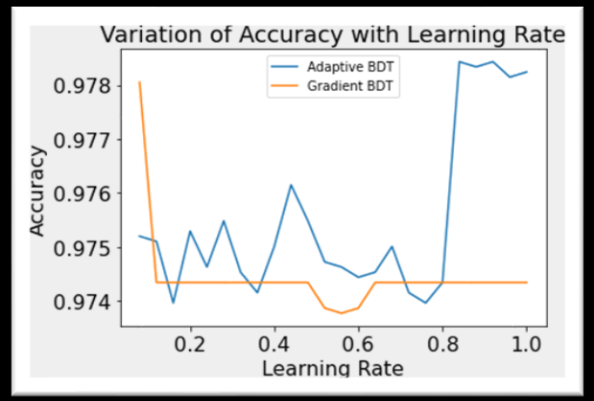


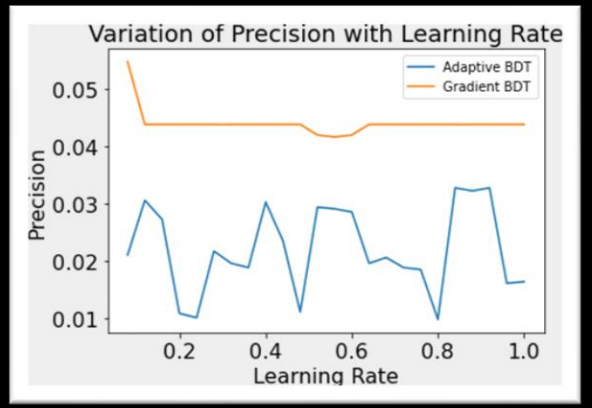


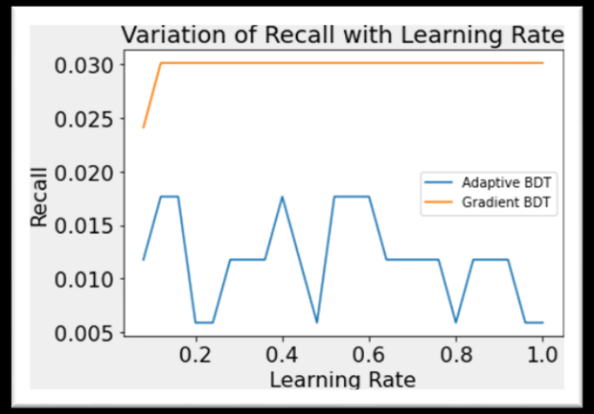


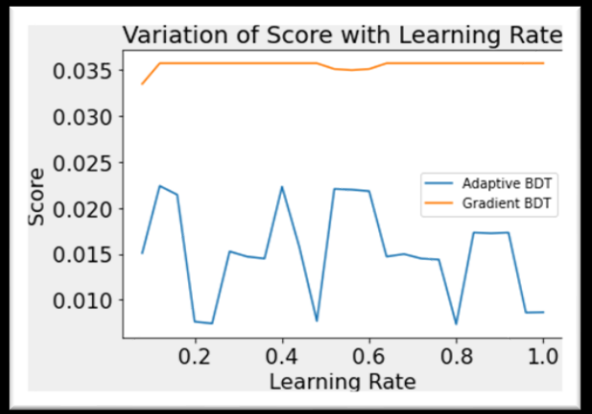


Comparison between Gradient Boosted and Adaptive Boosted Decision Trees









Comparison between Gradient Boosted and Adaptive Boosted Decision Trees with larger background dataset

Conclusion and future prospects

- A comparative analysis between Gradient Boosted Decision Tree (GBDT) and its Adaptive Boosted counterpart was done.
- The analysis showed that GBDT performed at the same level as Adaptive Boosted BDT.
- To further improve GBDT's performance, the article suggests exploring stochastic-based modeling or using Gradient Boosted Distributed Decision Tree.
- The analysis also suggests obtaining the di-electron invariant mass distribution for each classified event and estimating the branching fraction using the improvised code.

- Redundant features were manually removed, and the article suggests using an autoencoder-based unsupervised learning algorithm to further improve the process.
- The prepared BDT classifier's performance can be compared with neural network-based algorithms such as Recurrent Neural Network (RNN).
- The comparison can include the time complexity of the classifier algorithms for a more comprehensive distinction.

THANK YOU!