An analysis with Machine Learning in a nutshell

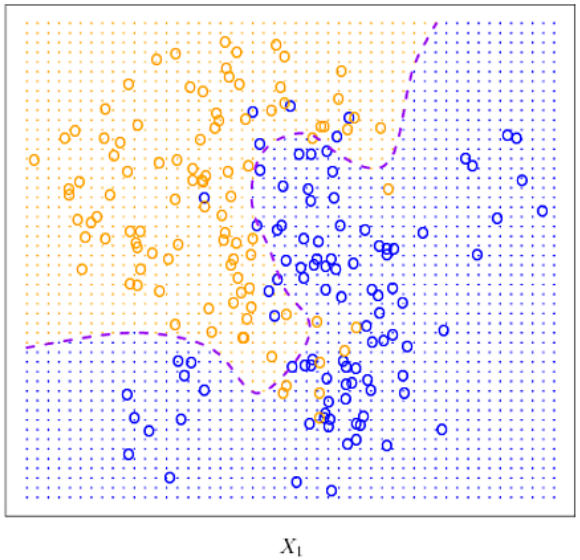
ALICE Starterkit 2019

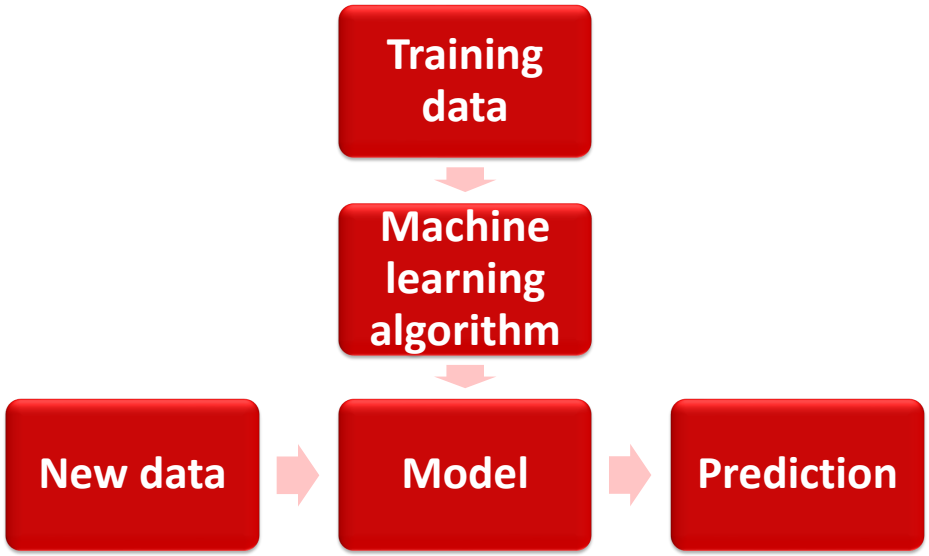
Fabio Catalano, Pietro Fecchio, Fabrizio Grosa, Luuk Vermunt

25 October 2019

Supervised machine learning

- Supervised Machine Learning (ML) models "learn" to make predictions from a set of examples, where the correct classification is known





- They can perform non-linear and more complex selections with respect to the linear selections traditionally used in particle physics

Would a ML model improve the analysis results?

Supervised machine learning

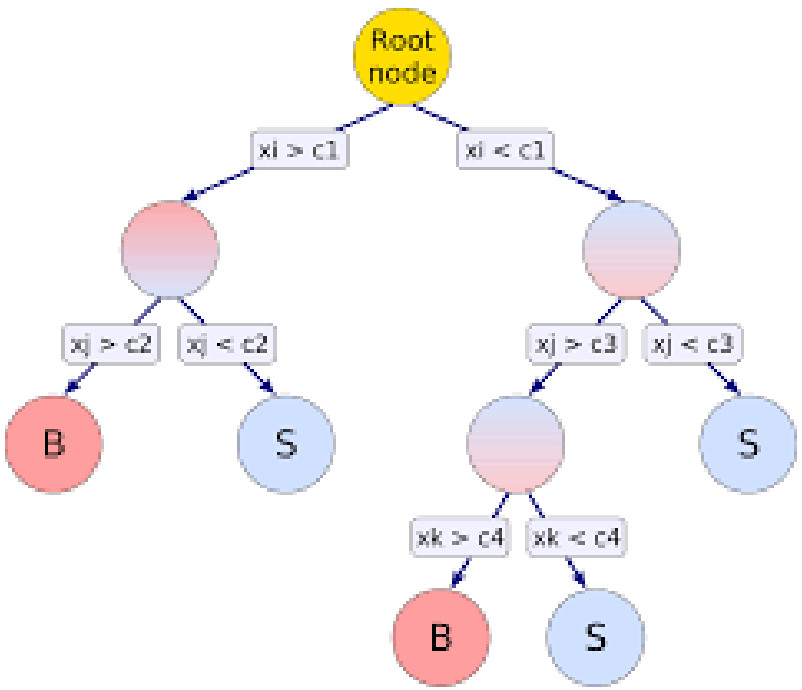
- To train the model, a training set which contains the examples is needed
- A test set is used to estimate the model performance
- Model predictions on real data → discrimination of the signal from the combinatorial background

Training and test set composed of:

- Signal candidates → **MC productions**
- Background candidates → **Data collected by the experiment** (sidebands of invariant-mass distribution, like-sign, event mixing, . . .)

Boosted Decision Trees (BDT)

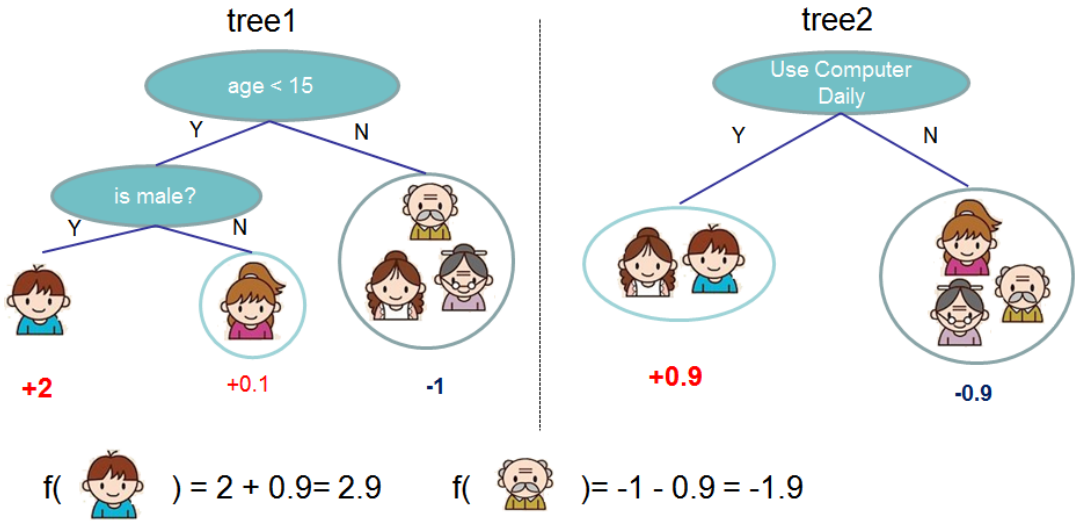
- Model: BDT. Effective in binary classification problems with high-level features and relatively few data



BDT are based on very simple decision trees:

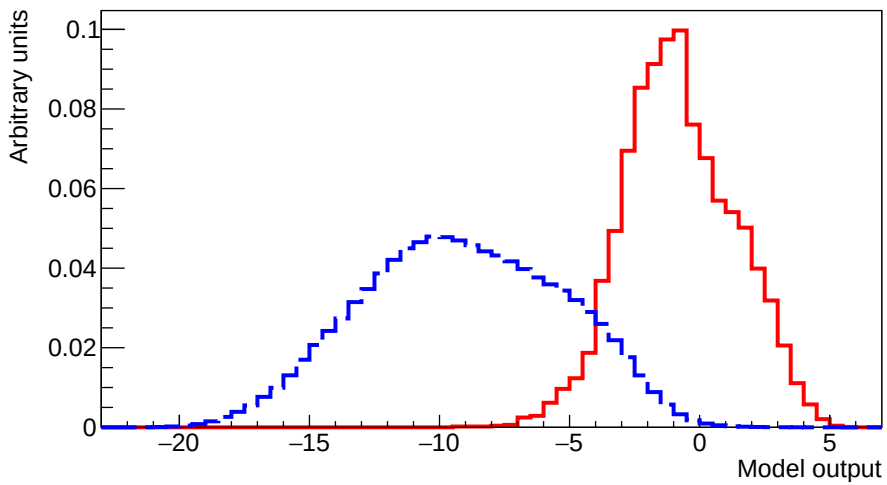
- The tree is built recursively utilizing the training set
- At each node the variable and its value that maximize the separation between signal and background is selected
- To quantify the goodness of the separation a score is defined (Gini index, entropy, . . .)

Boosted Decision Trees (BDT)



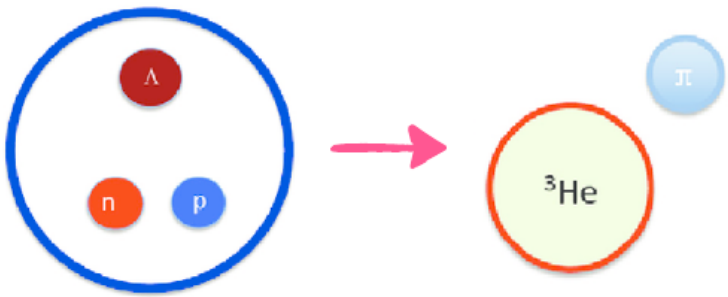
- A single tree has a poor prediction power ➔ combine $\mathcal{O}(100)$ to obtain a better model (*boosting*)
- XGBoost is based on a procedure called *gradient boosting* (in some ways similar to what is done in the neural network training)

- The BDT output depends on the candidate characteristics
- A *threshold* must be chosen to discriminate signal from background



Hypertriton in ALICE

- Crucial information on QGP provided by nuclei and hypernuclei ➔ sensitive to late stage of the system



Hypertriton

- lightest known hypernucleus, bound state of p, n, and Λ
- mass $\simeq 2.992\ \mathrm{GeV}/c^2$
- short-lived particle ➔ decays weakly in some hundred of ps

Decay channel

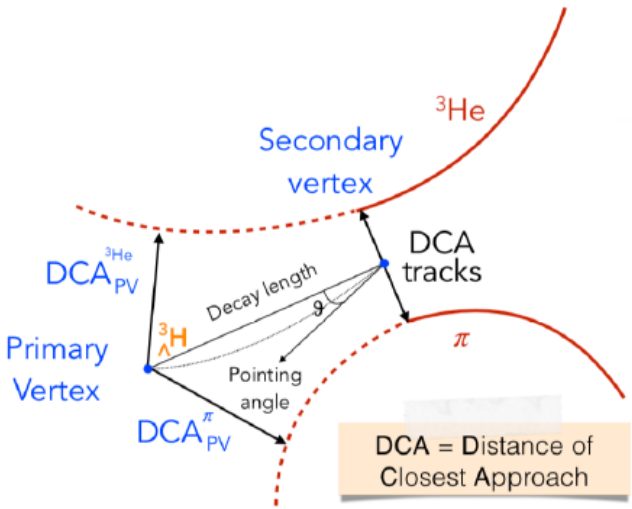
- mesonic
- non mesonic

The hypertriton can be measured via different charged mesonic decay

Decay
${}^{3}_{\Lambda}H \rightarrow {}^{3}He + \pi^{-}$
${}^{3}_{\Lambda}H \rightarrow d + p + \pi^{-}$

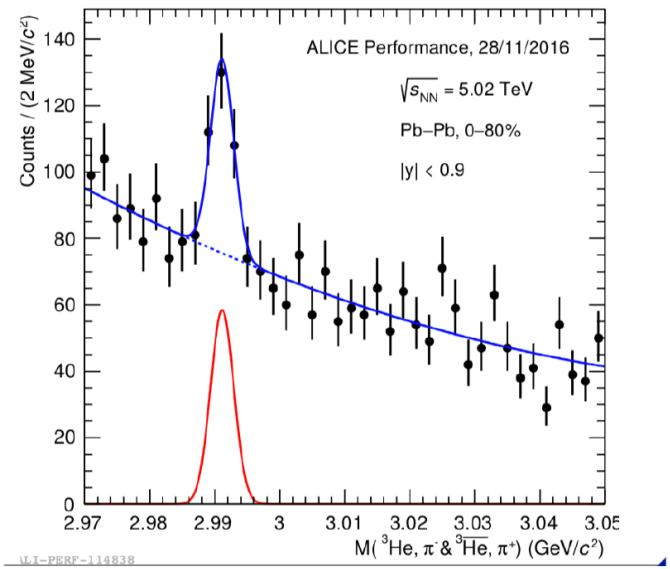
Hypertriton in ALICE

Two body decay ${}^{3}_{\Lambda}H \rightarrow {}^{3}He + \pi^{-}$



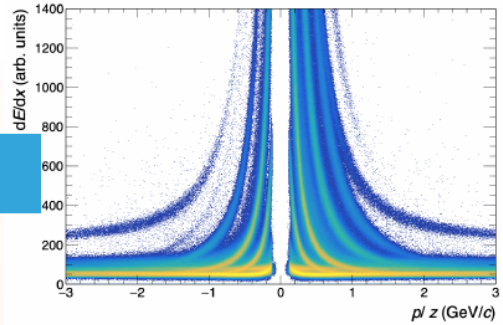
- Signal extracted with a fit to the invariant mass distribution
  - hypertritons are found applying selections on physical quantities measured by detectors
  - Is possible to improve these selections and get a better measurement?

- Candidates built from couple of tracks reconstructed at mid-rapidity with proper charge combination

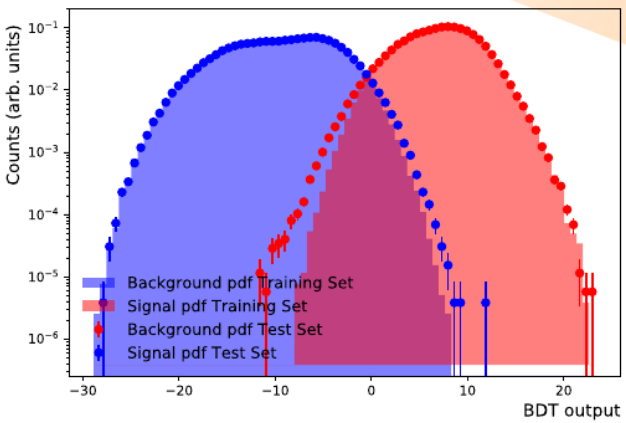


Recap — Data analysis flow

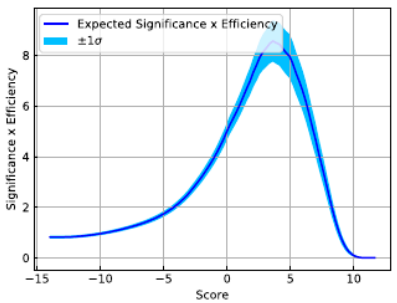
DATA PREPARATION

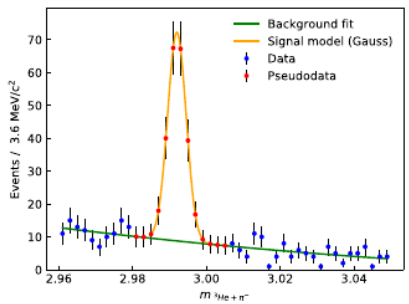


TRAINING AND TESTING THE MODELS

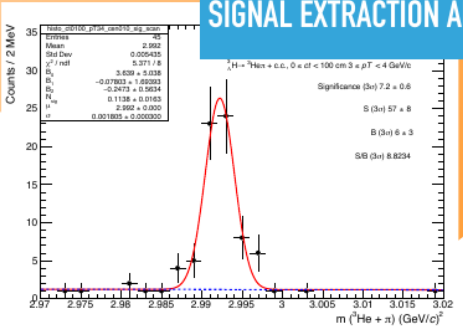


SELECTION OPTIMISATION





SIGNAL EXTRACTION AND MEASUREMENT



Let's start the tutorial

To start the tutorial go to this repository

https://github.com/fcatalan92/starterkitML19