Análisis de Datos con ROOT para HEP

Omar Zapata

Grupo de Fenomenología de Interacciones Fundamentales

Universidad de Antioquia

Centro de Ciencias de la Computación

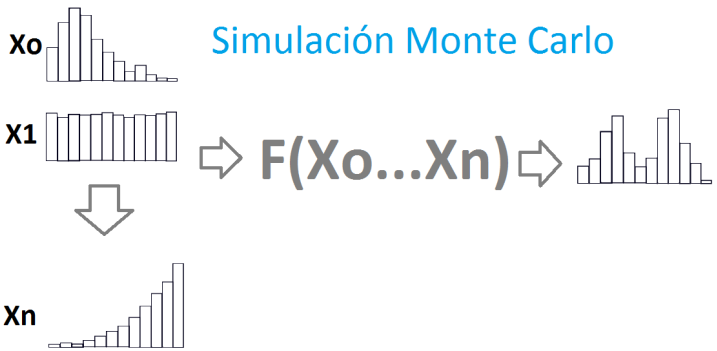
Instituto Tecnológico Metropolitano

Temas

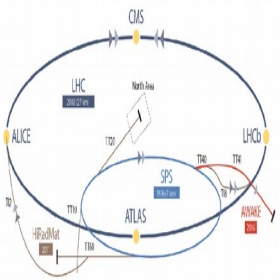
ROOT
Data Analysis Framework

- Overview del Análisis de Datos en HEP
- Data Mining
- Machine Learning
  - Redes neuronales
  - Árboles de decisión
  - Máquinas de soporte Vectorial
  - Clustering
- ROOT
- TMVA
- Python/R en ROOT

Overview
Análisis de Datos en HEP



Datos de los Aceleradores

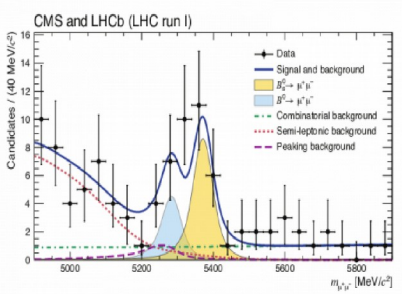


Machine Learning

Estadística

Procesamiento de Datos

Resultados



Data Mining

Obtener información importante de grandes volúmenes de datos.

Metodologías

- Bases de Datos
- Estadística
- Inteligencia Artificial (Machine Learning)

Juntas forman un área llamada Data Science, que estudia metodologías para el procesado(algoritmos nuevos y paralelización de código) y visualización de los datos.

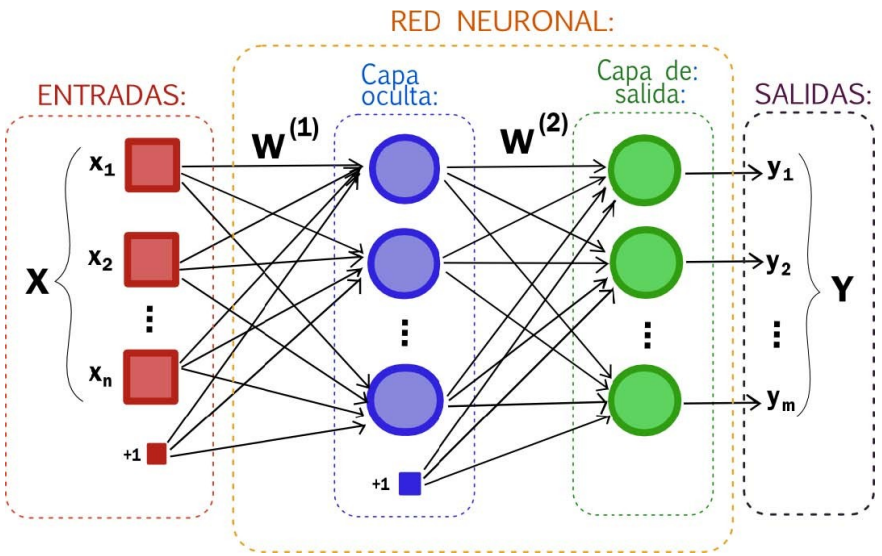
Machine Learning

En computer science machine learning se define como una rama de la inteligencia artificial encargada de desarrollar algoritmos capaces de identificar patrones y se usa principalmente para clasificación y regresión.

- Aprendizaje supervisado
  - Redes neuronales
  - Árboles de decisión
  - Máquinas de soporte vectorial
- Aprendizaje no supervisado
  - Clustering o agrupación.
- Algoritmos genéticos que son de los dos tipos.

Redes Neuronales

- Capas(layers)
- Las capas tienen diferentes tipos de funciones:
  - Inicialización
  - Propagación(Excitación)
  - Activación
  - Transferencia
  - Actualización
  - Pruning
- Tipos de redes:
  - RNA(Redes de Kohonen)
  - MLP(Multi-Layer Perceptron)
  - ART (Adaptative Resonance Theory)
  - Jordan
  - Etc..

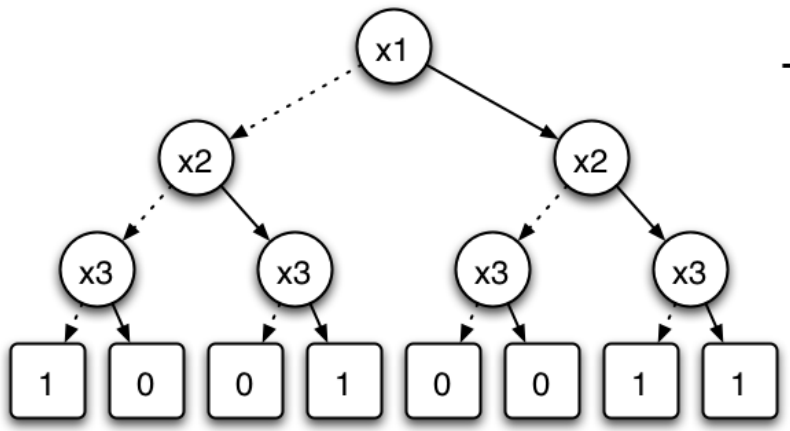


Árboles de Decisión

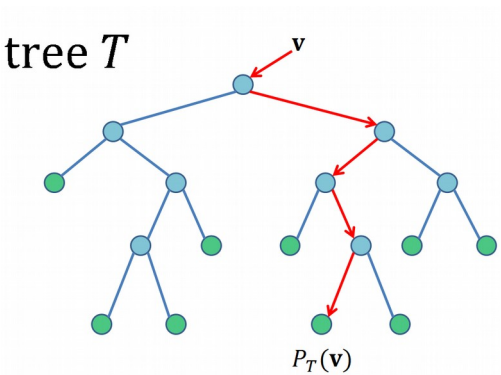
Son árboles binarios construidos a partir de un set de información dada.

Principales Métodos

- XGB (eXtreme Gradient Boost)
- Random Forest
- C50 (Information entropy)
- AdaBoost (Adaptative)

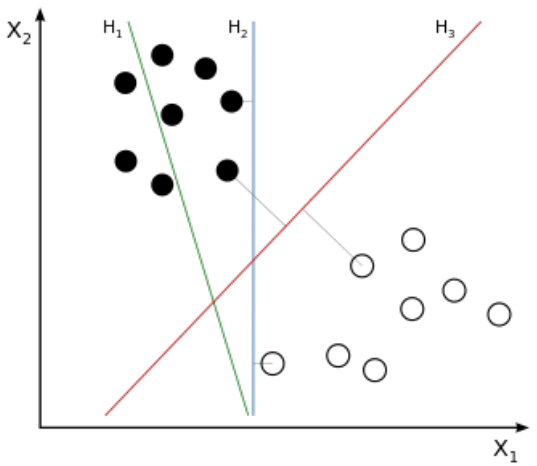


x1	x2	x3	f
0	0	0	1
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	0
1	1	0	1
1	1	1	1

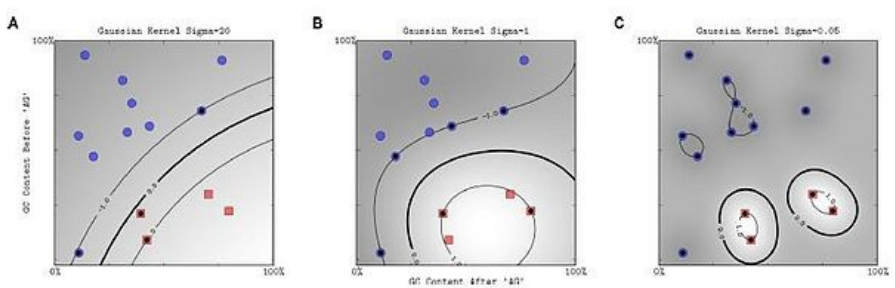


Máquinas de Soporte Vectorial

Algoritmo de agrupación que construye un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta (o incluso infinita) que fue desarrollado por Vladimir Vapnik y su equipo en los laboratorios AT&T.

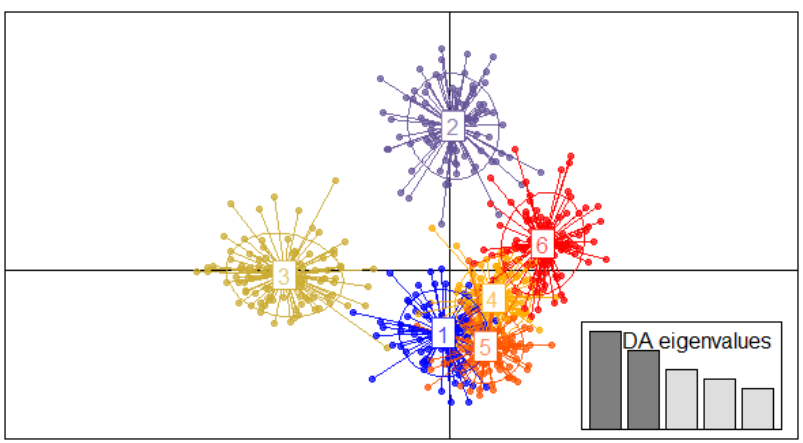


- Función Kernel (Delimita)
  - Ajuste de la función a los datos para agrupación.
  - Algunos kernels son:
    - Polinomial
    - Gaussiano
    - Perceptron



Clustering
Aprendizaje no supervisado

- Agrupación de variables categóricas
- Se lo considera una técnica de aprendizaje no supervisado puesto que busca encontrar relaciones entre variables descriptivas, pero no la que guardan con respecto a una variable objetivo.
- Algoritmos
  - Primeros vecinos
  - Lógica difusa
  - Algoritmos genéticos
  - Etc...



ROOT

Es un framework de análisis de datos científicos modular, que provee todas las funcionalidades necesarias para el procesamiento de grandes volúmenes de datos, análisis estadístico, visualización y almacenamiento de información. Esta escrito en C++ pero cuenta con integración a otros lenguajes de programación como Python y R.

Diseñado para HEP pero usado en muchas otras áreas como biología, química, astronomía etc..

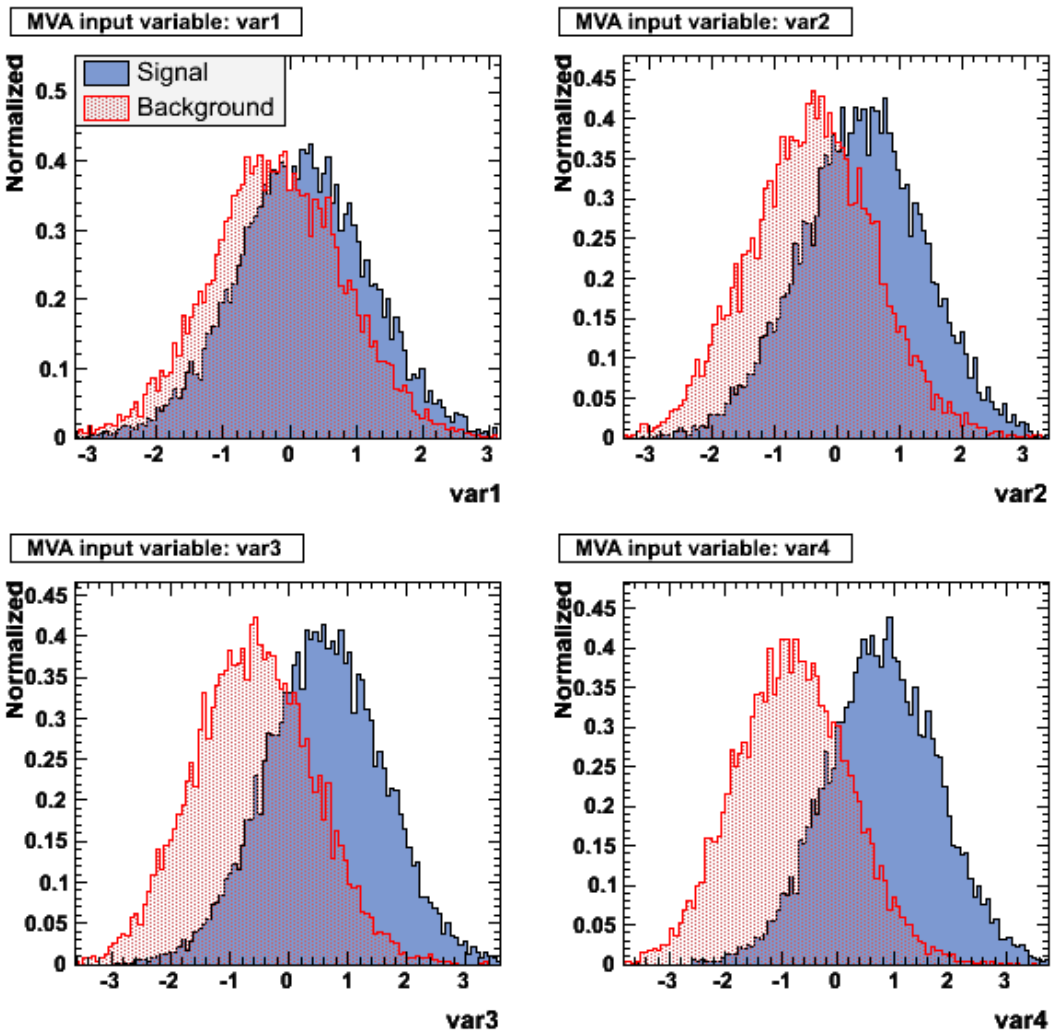
- Cuenta con un intérprete de línea de comandos.
- Un kernel para el proyecto Jupyter
- Visualización web con javascript JSROOT

TMVA

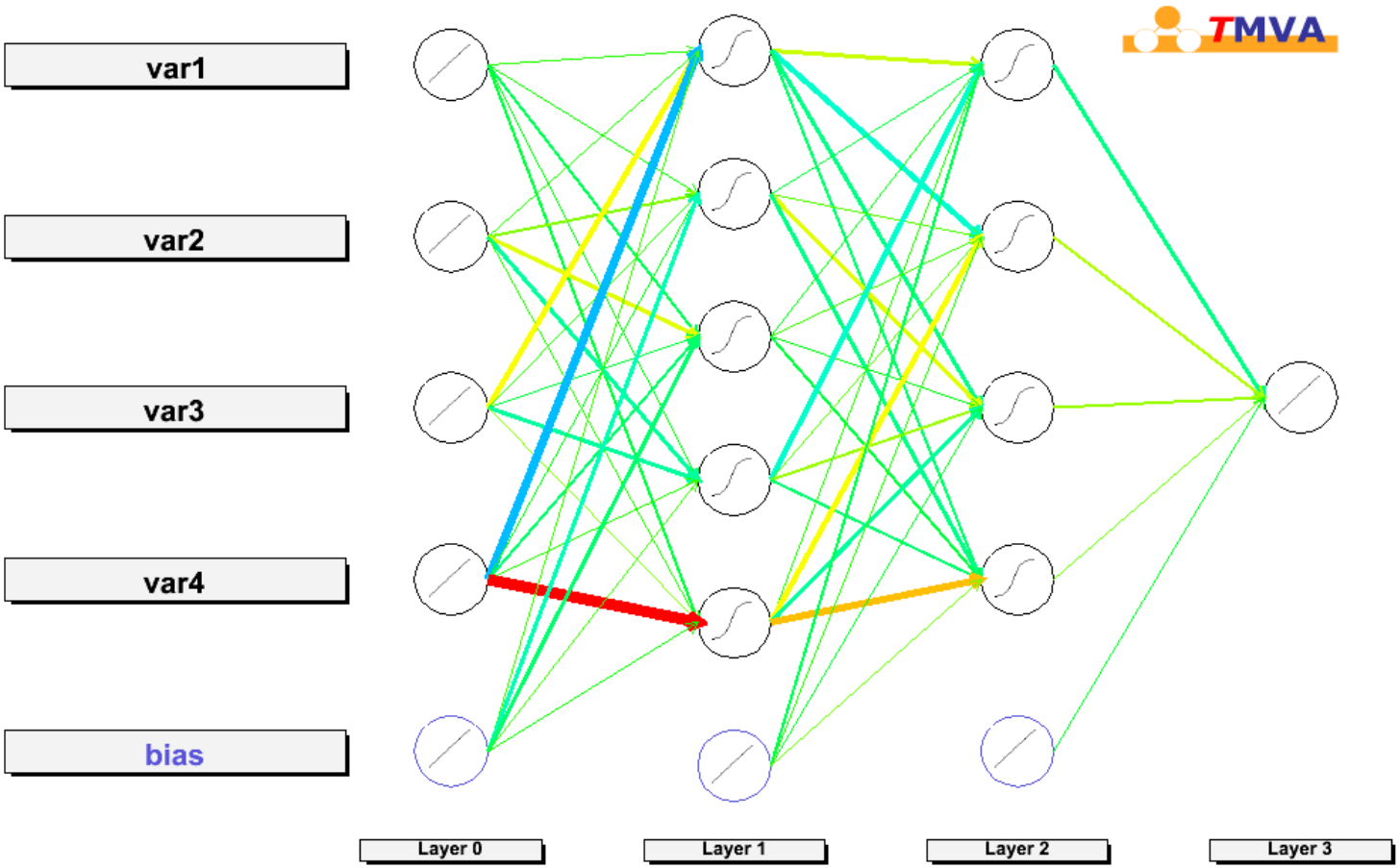
Toolkit for Multivariate Data Analysis (TMVA) es un paquete de ROOT para Machine Learning.

- Rectangular cut optimization
- Projective likelihood estimation (PDE approach)
- Multidimensional probability density estimation (PDE - range-search approach)
- Multidimensional k-nearest neighbour classifier
- Linear discriminant analysis (H-Matrix and Fisher discriminants)
- Function discriminant analysis (FDA)
- Artificial neural networks (three different implementations)
- Boosted/Bagged decision trees
- Predictive learning via rule ensembles (RuleFit)
- Support Vector Machine (SVM)

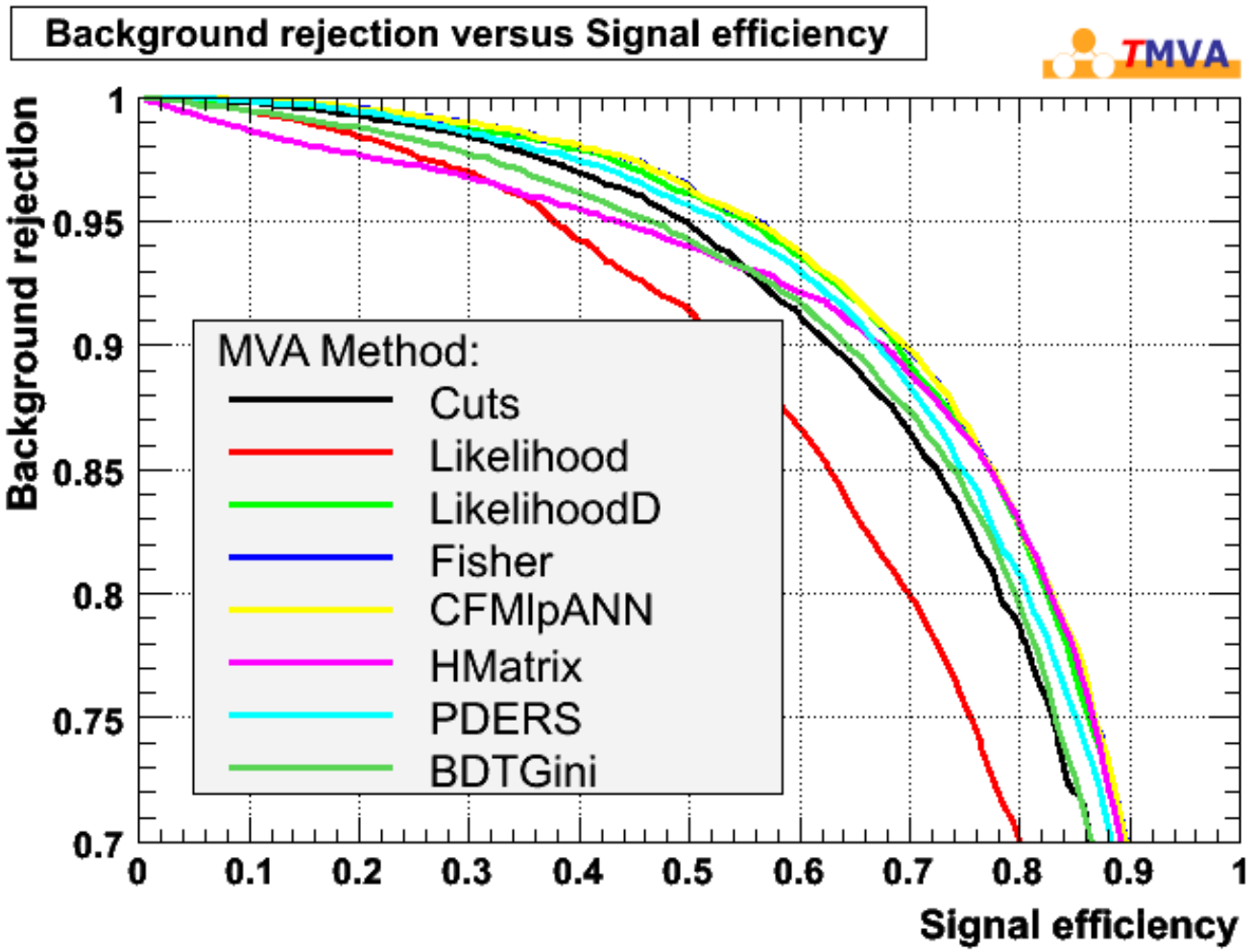
TMVA


MVA input variable: var1
Signal
Background
Normalized
var1
MVA input variable: var2
Normalized
var2
MVA input variable: var3
Normalized
var3
MVA input variable: var4
Normalized
var4

TMVA



TMVA


Background rejection versus Signal efficiency
TMVA
Background rejection
Signal efficiency
1
0.95
0.9
0.85
0.8
0.75
0.7
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
MVA Method:
Cuts
Likelihood
LikelihoodD
Fisher
CFMlpANN
HMatrix
PDERS
BDTGini

TMVA

¿Qué le falta?

Es una interfaz de hace 5 años por lo tanto:

- No tiene algoritmos nuevos
  - Deep Learning
  - Reduce SVM
  - Algoritmos genéticos
  - Etc..
- No esta paralelizado.
- No tiene utilidades para variable reduction.
- No tiene utilidades para cross validation.
- No es tan flexible para diseñar experimentos estadísticos como en R y Python.

TMVA new Desing

- **PyMVA (Python TMVA)**
  - **http://oproject.org/PyMVA**
- **RMVA (R TMVA)**
  - **http://oproject.org/RMVA**

  Variable Importance
- **Cross Validation**
- **Paralelización**

http://oproject.org/TMVA+Future

http://iml.cern.ch/

Referencias

- Imágenes
  - Decision Tree and Decision Forest - File Exchange - MATLAB Central. (n.d.). Retrieved February 22, 2016, from http://www.mathworks.com/matlabcentral/fileexchange/39110-decision-tree-and-decision-forest
  - Cern. (n.d.). Retrieved February 22, 2016, from http://eu-datagrid.web.cern.ch/eu-datagrid/
  - Coursera (n.d.). Retrieved February 22, 2016, from https://class.coursera.org/ml-005/lecture
  - Graph-Based Algorithms for Boolean Function Manipulation, Randal E. Bryant, 1986
  - Sharma, Ashok (2015). Retrieved February 22, 2016, from http://image.slidesharecdn.com/machinelearning-150304000629-conversion-gate01/95/machine-learning-18-638.jpg?cb=1425429333
  - Wikiwand (n.d.). Retrieved February 22, 2016, from http://www.wikiwand.com/en/Binary_decision_diagram

Referencias

- **Videos**
  - **https://cds.cern.ch/record/1541893?ln=en LHC Data**
  - **http://cds.cern.ch/record/2020780 CERN Overview**
  - **https://www.youtube.com/watch?v=qv6UVOQ0F44 Mario**

Thanks

Omar.Zapata@cern.ch

http://root.cern.ch
http://oproject.org