REDS - TP1 : Analyse préliminaire des données

Laura Nguyen et Keyvan Beroukhim

25 septembre 2019

La base de données issue du *ATLAS Higgs Boson Machine Learning Challenge 2014* est constituée de 818238 évènements simulés pouvant être soit des collisions *"Higgs to tautau"*, soit du *background*. Étant données ces deux classes, *"s"* pour *signal* et *"b"* pour *background*, l'objectif est de classifier au mieux les évènements.

	EventId	DER_mass_MMC	DER_mass_transverse_met_lep	DER_mass_vis	DER_pt_h	DER_deltaeta_jet_jet	DER_mass_jet_jet	DER_prodeta_jet_jet	DER_
407755	507755	168.108	90.494	78.229	50.782	-999.000	-999.000	-999.000	
652847	752847	63.332	67.308	57.416	2.740	-999.000	-999.000	-999.000	
239074	339074	118.435	66.054	54.824	108.381	2.730	195.210	4.842	
654458	754458	98.325	21.762	73.594	28.887	-999.000	-999.000	-999.000	
115891	215891	107.612	14.255	52.112	178.665	6.419	1762.835	-10.292	

Fig. 1: Échantillon de 5 évènements du dataset avec uniquement les 8 premiers attributs affichés

Chaque évènement est défini par 35 attributs, dont son label. Nous retirons du dataset les features servant uniquement pour le challenge Kaggle : `Weight`, `KaggleSet` et `KaggleWeight`.

Les données sont composées à 34% de signaux et 66% de background. Nous remarquons cependant que le dataset contient, à vue d'œil, un nombre important d'exemples avec des valeurs manquantes (indiquées par -999). Or, travailler avec des datasets contenant des données manquantes est compliqué. Nous décidons donc de supprimer les évènements dont au moins un attribut n'est pas valide et nous nous retrouvons avec seulement 223574 exemples : restreindre le dataset aux données complètes fait perdre 75% des données. De plus, la base restreinte contient 47% de signaux, soit 13% de plus que dans le dataset original. Si ce dernier est représentatif de la réalité alors celui restreint ne l'est pas.

Les figures 2, 3 et 4 contiennent les heatmaps des matrices de corrélation associées à différentes versions du dataset.

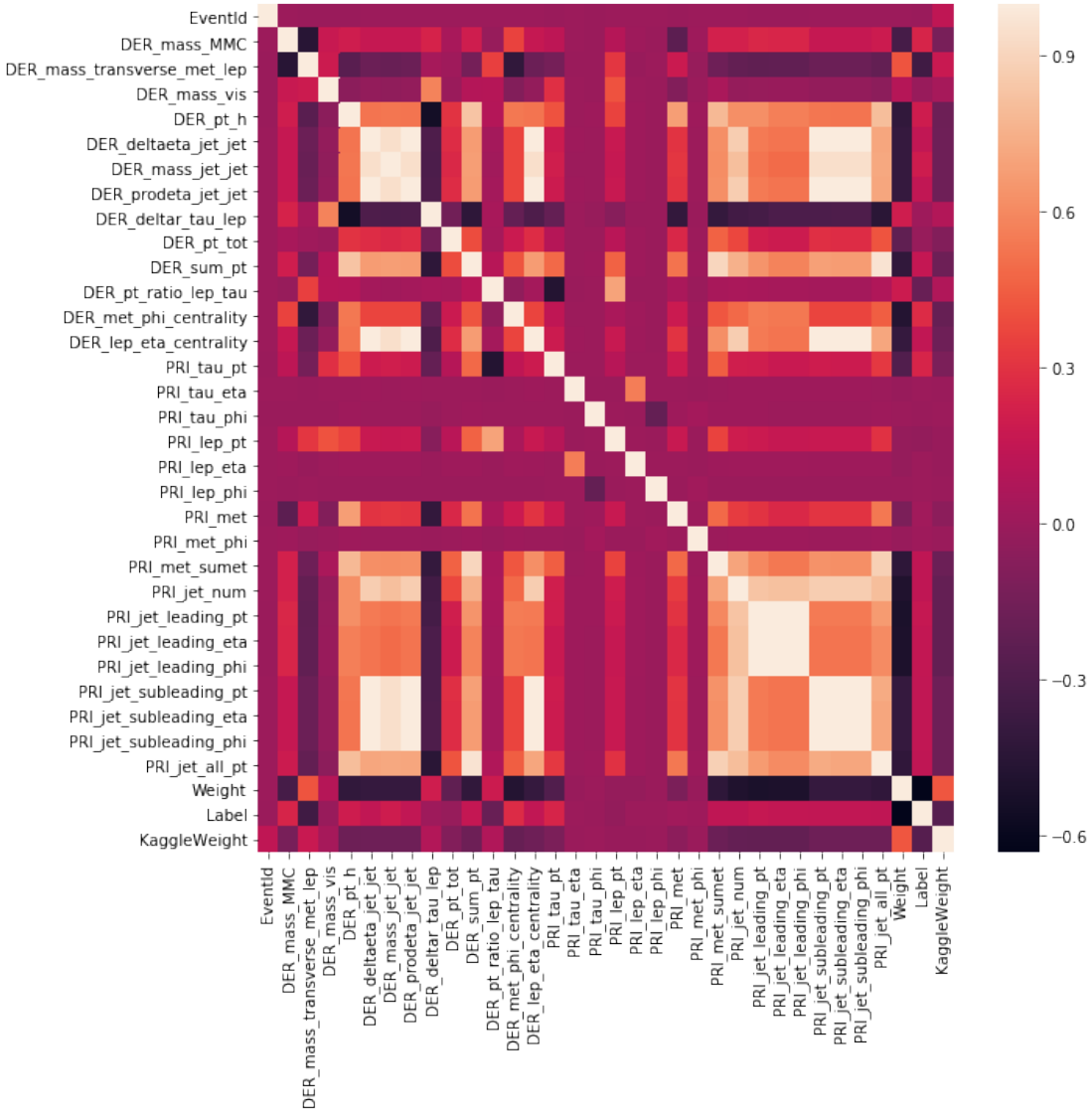

Fig. 2: Heatmap de la matrice de corrélation avec conservation des valeurs invalides

La heatmap associée à la matrice de corrélation entre chaque attribut du dataset original est présentée sur la figure 2.

La variable `EventId` ne fournit pas d'information : cela se remarque, par exemple, par une corrélation nulle avec toutes les autres variables. Par la suite, nous ne considérerons donc pas cette variable.

Certaines variables sont fortement corrélées, d'autres anti-corrélées, e.g. `DER_mass_MMC` et `DER_mass_transverse_met_lep`. Ces informations pourrnt se révéler utiles si nous souhaitons traiter les évènements à données manquantes.

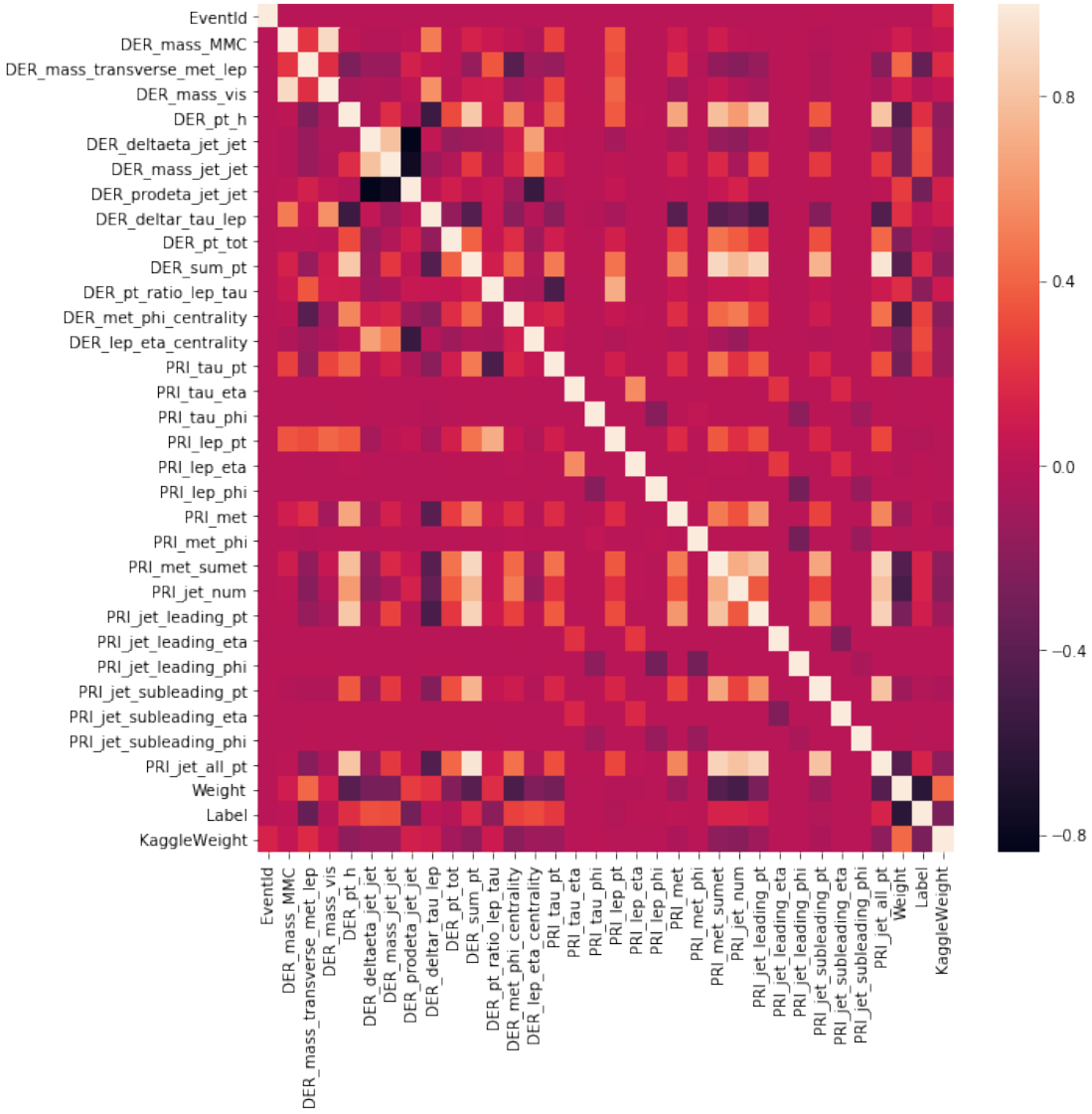

Fig. 3: Heatmap de la matrice de corrélation en ignorant les valeurs invalides

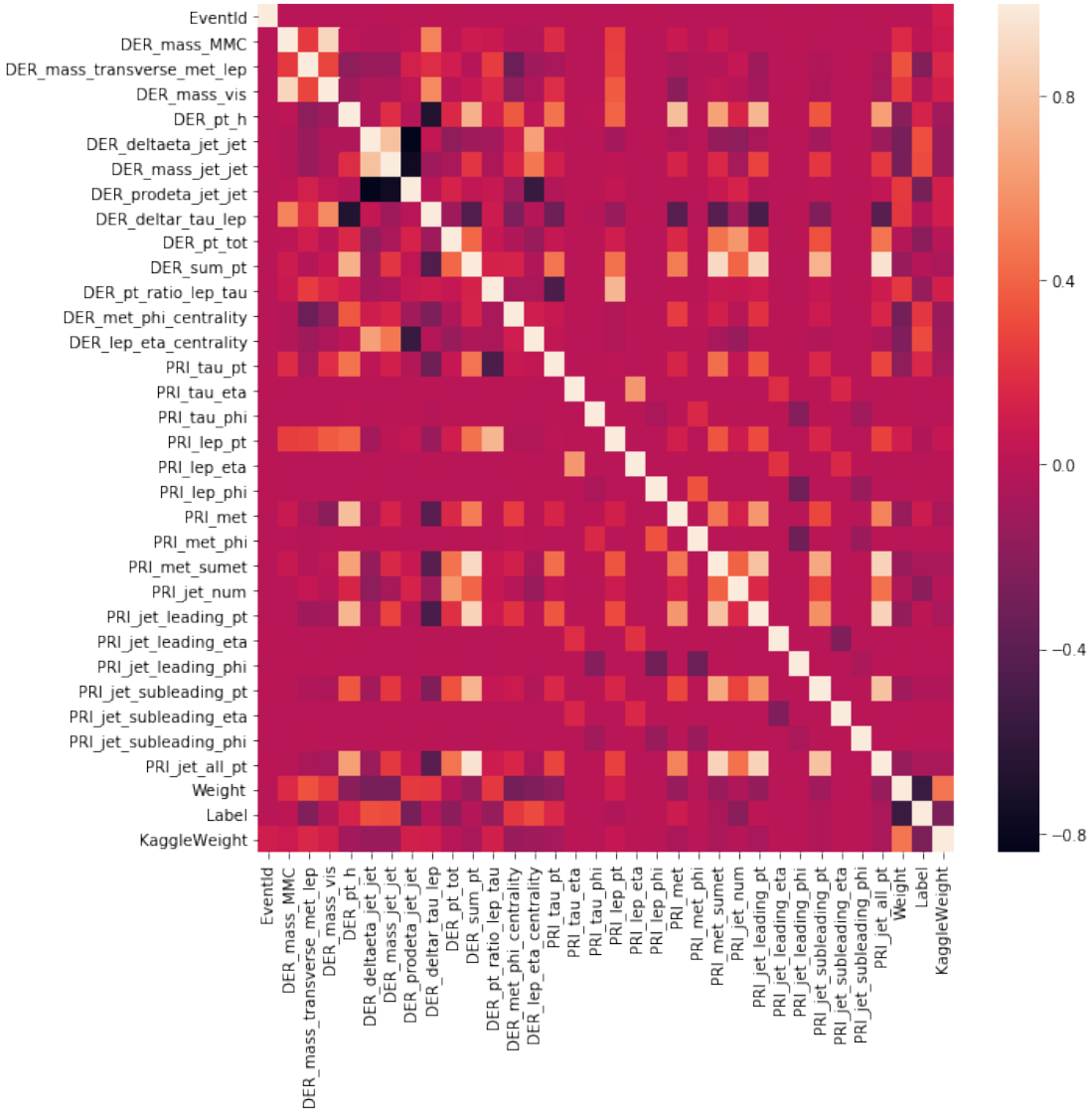

Fig. 4: Heatmap de la matrice de corrélation en ignorant les évènements contenant des valeurs invalides

D'après les figures 3 et 4, les heatmaps de la matrice de corrélation obtenue en ignorant les valeurs invalides et celle obtenue en ignorant les évènements ayant au moins un attribut à valeur invalide sont très similaires.

	coefficient de corrélation
PRI_tau_phi	-0.000138846228099
EventId	-0.000504580072614
PRI_jet_leading_eta	0.000788635648117
PRI_jet_leading_phi	-0.00137544407398
PRI_jet_subleading_eta	0.00185630137404
PRI_lep_eta	-0.00428147335566
PRI_lep_phi	0.00453211325684
PRI_tau_eta	-0.00475718760362
PRI_jet_subleading_phi	-0.00595329161404
PRI_met_phi	0.00608780781447
PRI_jet_leading_pt	0.0186964320308
DER_mass_MMC	0.0206623199582
DER_sum_pt	-0.0230014614961
DER_deltar_tau_lep	-0.023584025589
DER_mass_vis	-0.0295019957663
PRI_jet_subleading_pt	-0.0381650708107
PRI_lep_pt	-0.0457516673814
PRI_jet_all_pt	-0.058394877924
PRI_met_sumet	-0.0658624099293
PRI_met	0.0860522440004
DER_pt_h	0.133548075576
DER_pt_ratio_lep_tau	-0.146528336278
PRI_tau_pt	0.157097764577
PRI_jet_num	-0.19434643912
DER_pt_tot	-0.194625620451
DER_met_phi_centrality	0.224938504456
DER_mass_transverse_met_lep	-0.235461340206
KaggleWeight	-0.261312207528
DER_prodeta_jet_jet	-0.286358935457
DER_mass_jet_jet	0.303816127333
DER_lep_eta_centrality	0.30496119586
DER_deltaeta_jet_jet	0.319156166812
Weight	-0.554073494618
Label	1.0

Fig. 5: Coefficients de corrélation entre chaque attribut et la classe `Label`

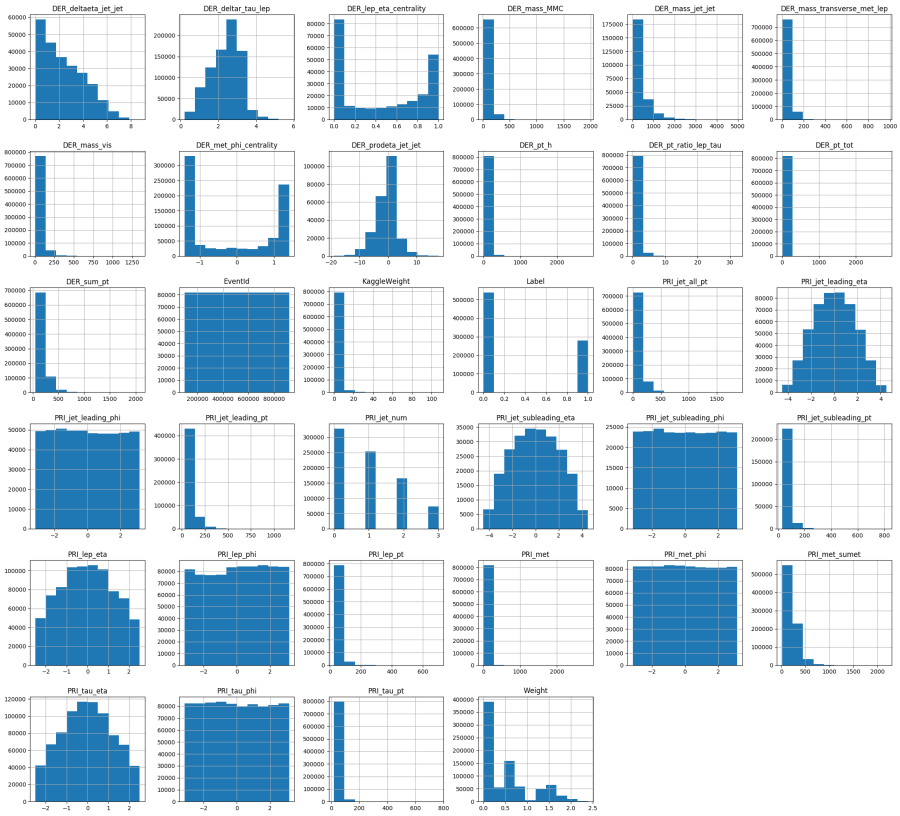

Fig. 6: Histogramme des valeurs de chaque attribut du dataset où les valeurs manquantes sont ignorées

La figure 6 contient les histogrammes de valeurs pour chaque attribut du dataset où les valeurs manquantes sont ignorées. Plusieurs histogrammes sont très déséquilibrés, cela est peut-être dû à une présence d'erreur dans les données. Nous pourrions essayer de traiter ces cas-là en les supprimant par exemple.