REDS - Projet Boson de Higgs

Rapport

Kim-Anh Laura Nguyen
Keyvan Beroukhim
Master 2 DAC
Promo 2019-2020

Enseignant : Olivier Schwander

1 Introduction

Nous souhaitons détecter la présence du boson de Higgs dans des données simulées dans le but de reproduire le comportement de l'expérience ATLAS. Il s'agit d'un problème de détection d'évènement, ou classification binaire, dans lequel les deux classes sont :

— *background*

— *tau tau decay of a Higgs boson*

Les données sont issues du projet Kaggle *ATLAS Higgs Boson Machine Learning Challenge 2014*.

2 Analyse préliminaire des données

La base de données est constituée de 818238 évènements simulés. Chaque évènement est défini par 30 attributs numériques et un label à prédire.

Les données sont composées à **34% de labels positifs** (les *signaux*) et à 66% de labels négatifs (le *background*). Les classes ne sont pas suffisamment déséquilibrées pour gêner l'entraînement. Par ailleurs, la métrique propre à cette tâche, nommée *AMS*, nous est imposée. Le choix habituel de la mesure d'évaluation à utiliser pour prendre en compte ce déséquilibre ne se pose donc pas.

2.1 Distribution des variables

La figure 1 contient les histogrammes de répartition des valeurs de quelques variables, en omettant les valeurs manquantes. Nous constatons que **toutes les variables ne suivent pas le même type de distribution**. Or, les algorithmes d'apprentissage se comportent généralement mieux avec une distribution des données équilibrée. Dans ce dataset, de nombreuses variables ont une distribution exponentielle décroissante (e.g `DER_MASS_VIS` sur la figure 1).

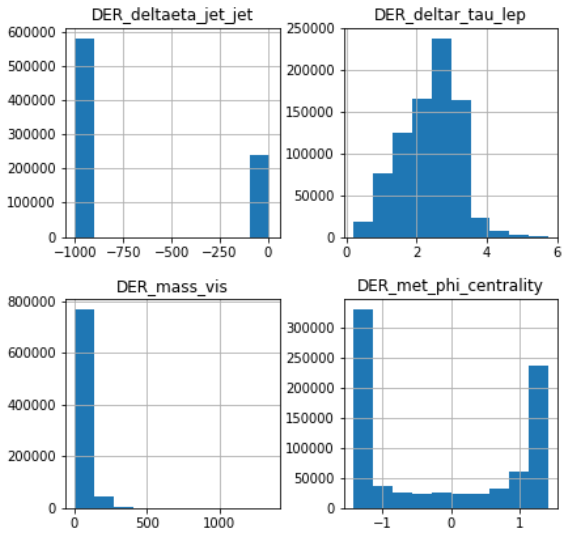
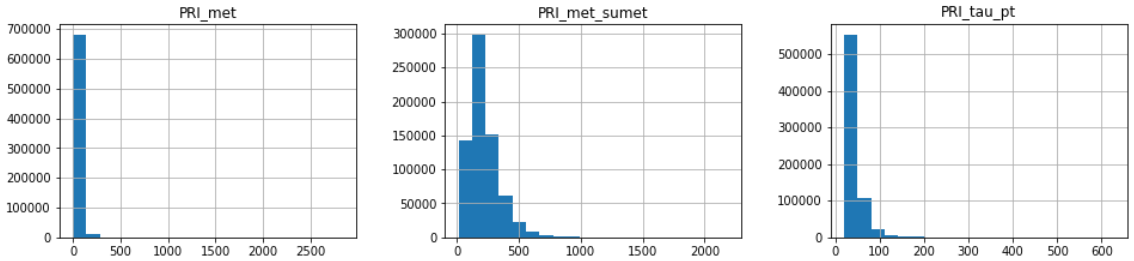
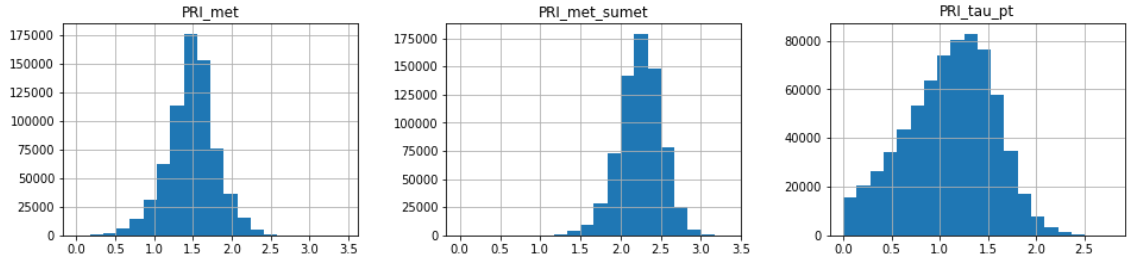


FIGURE 1 – Histogrammes de répartition des valeurs de quelques variables

Afin d'obtenir des **distributions normales** et de **stabiliser la variance**, nous effectuons une **log-transformation** de manière indépendante sur les colonnes. Nous commençons par ajouter une constante à chaque colonne à transformer pour que la valeur minimale de sur cette colonne soit 1 (comme le log ne prend que des valeurs strictement positives). Nous appliquons ensuite la fonction log sur les colonnes. Avec ce dataset et en utilisant cette méthode, nous obtenons toujours des distribution normales. Finalement nous **centrons et réduisons** chaque variable de manière indépendante car cela facilite généralement l'apprentissage des modèles.



(a) Avant log-transformation



(b) Après log-transformation

FIGURE 2 – Histogrammes de répartition des valeurs de quelques variables avant et après log-transformation

2.2 Valeurs manquantes

Le dataset contient **21% de valeurs manquantes** (indiquées par la valeur -999). D'une manière générale, les modèles d'apprentissage statistique ne gèrent pas automatiquement les valeurs manquantes. Il faut donc traiter ces données avant de les présenter à nos modèles. Nous considérons trois façons de pallier à ce problème.

2.2.1 Omission

En supprimant les évènements dont au moins un attribut n'est pas valide, nous retrouvons avec seulement 223574 exemples d'apprentissage : **restreindre le dataset aux données complètes fait perdre 75% des évènements**. Par ailleurs, la base restreinte contient 47% de labels positifs, soit 13% de plus que le dataset original. Cela met en évidence le fait que les données ne sont pas manquantes de manière indépendante : l'absence des données est liée à leur valeur (nous sommes dans un contexte *Missing Not At Random*). Un modèle entraîné sur le dataset restreint en faisant l'hypothèse "i.i.d." serait donc biaisé.

2.2.2 Omission d'attributs

Une autre méthode consiste à **supprimer les features dont le pourcentage de valeurs manquantes dépasse un certain seuil**. On **retire ensuite les évènements à valeurs manquantes**. Le dataset final ne contient donc plus aucune donnée manquante. En fixant le seuil à 40%, notre jeu de données passe de 818238 à 693636 échantillons, soit 85% des données initiales, et de 35 à 20 attributs.

2.2.3 Conservation

Les ***Random Forest*** font partie des algorithmes permettant de **gérer les données manquantes**. En effet, les arbres de décision peuvent **reconnaître une donnée manquante par un simple test de valeur**.

2.2.4 Affectation (*Imputing*)

Une autre méthode consiste à **compléter les données manquantes** de la manière la plus "intelligente" possible. Une fois les données complétées, cela permet d'**utiliser tous les modèles de classification à notre disposition**.

Nous nous servons de l'algorithme `IterativeImputing` de scikit-learn, qui complète de manière itérative les données manquantes en utilisant des modèles de régression prédisant la valeur d'une variable à partir de la valeur des autres variables.

En pratique et comme nous souhaitons conserver le maximum d'information, **nous ne supprimerons ni d'évènements ni d'attributs à valeurs manquantes**. Nous testerons, d'un côté, des **algorithmes permettant de gérer les données manquantes** et, de l'autre, la méthode de **complétion des données**. Pour cette dernière, nous appliquons d'abord les **log-transformations** puis nous réalisons l'**imputing** avant de finir par **scaler** les données. Ce choix se justifie premièrement par l'intuition que l'imputing fonctionnera mieux sur des données log-transformées et, deuxièmement, par le fait qu'estimer la moyenne et la variance en ignorant les données manquantes serait biaisé.

2.3 Sélection de features

Les données sont représentées par **30 variables explicatives**. Ce nombre n'étant **pas très élevé** (par exemple comparé au nombre de pixels dans une image), il n'y a donc **pas de grand risque de sur-apprentissage** pour les modèles. Ainsi, la sélection de caractéristique ne semble pas nécessaire. Cependant, comme le montre la figure 3, certaines variables sont **fortement corrélées** (e.g. `DER_mass_MMC` et `DER_mass_transverse_met_lep`). Or, des attributs corrélés peuvent diminuer la performance de certains algorithmes (en contribuant au sur-apprentissage par exemple).

Afin de réduire le nombre de variables explicatives, nous comparerons donc les performances avec ou sans l'utilisation d'une ***Analyse en Composantes Principales*** **(PCA)**. Pour **choisir le nombre de composantes principales**, nous traçons le **pourcentage de variance expliquée en fonction du nombre de composantes** (figure 4). Avec **20 composantes**, nous conservons 98.2% de la variance totale.

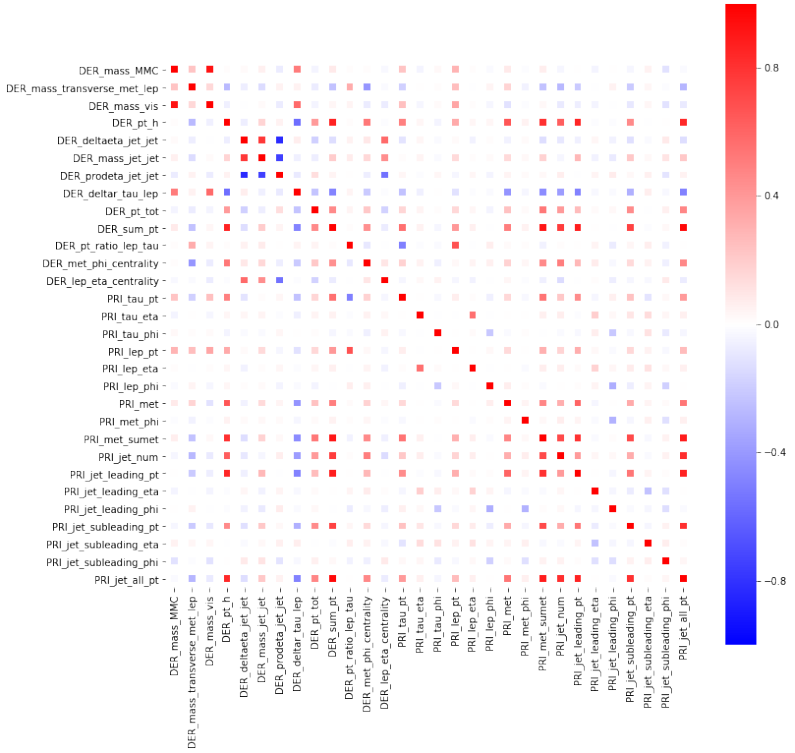


FIGURE 3 – Matrice de corrélation entre chaque attribut

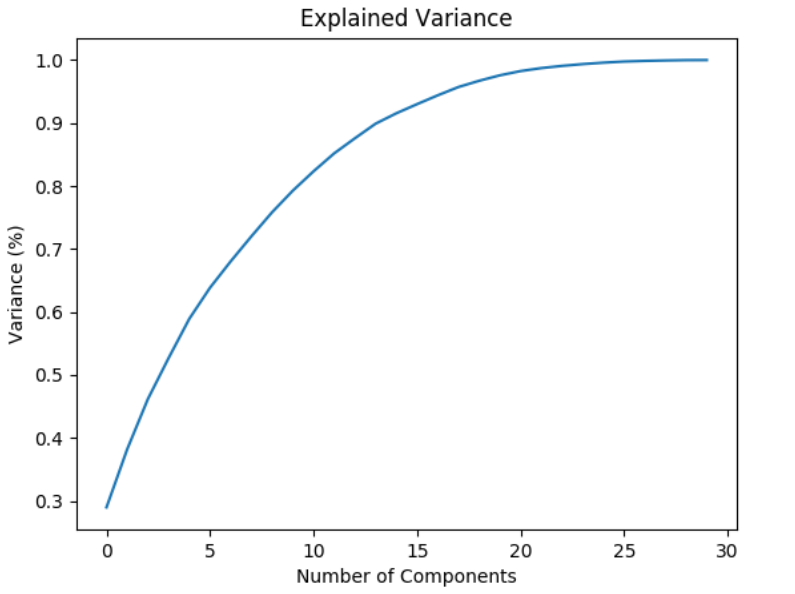


FIGURE 4 – Pourcentage de variance expliquée en fonction du nombre de composantes

3 Classification des évènements

3.1 Modèles utilisés

Nous considérons plusieurs modèles d'apprentissage pour classifier les évènements :

— SVM
— méthodes ensemblistes : Bagging de Perceptron, Random Forest, AdaBoost

3.1.1 Support Vector Machines

Les *Support Vector Machines* (SVMs) nous fournissent une ***baseline***. Généralisation des classifieurs linéaires, ils nécessitent un **faible nombre d'hyperpararamètres**, ont des **garanties théoriques** et donnent de **bons résultats** en pratique.

3.1.2 Méthodes d'ensemble

Afin de **réduire de réduire le biais et la variance** de nos modèles et ainsi améliorer les performances, nous utilisons des **méthodes d'ensemble**. Ces algorithmes consistent à **combiner plusieurs classifieurs faibles** pour obtenir un modèle **plus stable, plus robuste et ayant une meilleure capacité de généralisation**.

— Bootstrap AGGregatING (Bagging) :
Le *Bootstrap Aggregating* est une méthode pour **réduire la variance par moyennage**. Plusieurs ensembles d'apprentissage sont simulés par ***bootstrap*** **(tirage avec remise)** puis **sur chaque sous ensemble, on entraîne un classifieur faible**. Dans notre contexte de classification binaire, la prédiction du modèle est définie par le **vote majoritaire des modèles simples (*aggregating*)** .
Comme chaque classifieur faible est entraîné sur des données légèrement différentes, le modèle d'ensemble est capable de **capturer de petites variations dans les données** et donc de mieux **généraliser**.
Nous choisissons, uniquement pour le Bagging, le **Perceptron** comme classifieur faible. Il s'agit d'un **modèle instable**, ce qui permet d'obtenir un **ensemble de classifieurs suffisamment différents**.

— Random Forest :
Avec la technique du Bagging, même si chaque classifieur est appris sur un jeu de données légèrement différent, tous les modèles se servent des mêmes attributs pour partitionner les données, les arbres générés risquent donc d'être **trop similaires**. Pour pallier à ce problème, les arbres des **forêts aléatoires (*random forests*)**, également appris sur des bootstrap de l'ensemble original, **sélectionnent pour chacun de leurs noeuds un échantillon aléatoire de features**, ce qui **force les modèles à être différents** et empêche le modèle global de se concentrer sur des attributs particuliers.

— AdaBoost :
Avec la technique de Bagging, les classifieurs faibles sont appris indépendamment les uns des autres. Or, si ce classifieur est, à lui seul, trop faible, **le Bagging ne permettra pas d'obtenir un meilleur biais**. Le ***Boosting*** permet de traiter ce problème par **apprentissage successif de classifieurs faibles** : en donnant plus de poids aux exemples mal classés par les modèles précédents, **chaque nouveau classifieur se focalise sur les parties de l'espace mal prédites**. Contrairement au Bagging, la

contribution d'un classifieur à la prédiction du modèle d'ensemble est déterminée par sa performance.
En revanche, si le classifieur faible est trop complexe, le modèle de Boosting **risque de sur-apprendre**, auquel cas la technique de Bagging est plus adaptée.

3.2 Protocole d'expérimentation

Tout d'abord, nous commençons par **séparer le jeu de données en train et en test**. Le protocole d'expérimentation est le suivant :

a. Avant d'entraîner nos modèles, nous apprenons les **paramètres de prétraitement des données** sur l'ensemble d'apprentissage, et nous appliquons les prétraitements sur l'ensemble des données. Les prétraitements et apprentissages des paramètres sont effectués dans l'ordre décrit précédemment : **log-transformation** (valeurs minimales des colonnes apprises), **imputing** (paramètres de tous les régresseurs appris) et **scaling** (moyennes et écarts-types des colonnes appris).
b. Pour chaque algorithme, nous procédons par ***grid search* pour trouver les hyperparamètres optimaux**. Pour chaque combinaison d'hyperparamètres, la performance du modèle produit est estimée par ***cross-validation*** sur l'ensemble d'apprentissage.
c. Pour chaque modèle, nous récupérons les hyperparamètres ayant produit le meilleur résultat et **calculons le score final par cross-validation sur les données de test**.

Nous comparons les résultats obtenus à ceux produits en appliquant aux données, **en plus des prétraitements de l'étape a.**, une **PCA** (à 20 composantes principales).

En parallèle, pour pouvoir **comparer l'approche de complétion à celle de conservation des données**, nous menons d'autres expériences où les données sont prétraitées de la manière suivante : log-transformation, scaling et **remplacement des valeurs manquantes par une valeur aberrante**. Nous ne rajoutons pas de PCA à ces prétraitements.
Le seul modèle entraîné sur le jeu de données contenant des données aberrantes est un **Random Forest** que nous nommerons **RFnan**.

Par souci de temps de calcul, les expériences de recherche de paramètres optimaux et d'évaluation des modèles ne sont pas menées sur le dataset entier. Pour les **métriques usuelles**, **le score obtenu par un modèle est indépendant de la taille du jeu de test**, mais **ce n'est pas le cas pour la métrique AMS**. Afin d'obtenir des scores en AMS comparables, nous fixons la taille des jeux de tests à 1000 éléments.

Les scores obtenus en AMS étant difficilement interprétables, nous utilisons aussi comme métrique le **taux de bonne classification (*accuracy*)**. Pour chaque métrique, les scores présentés par la suite sont obtenus en optimisant les modèles pour cette métrique.

3.3 Résultats et analyse

Pour le **SVM**, nous cherchons la valeur optimale de la **pénalité C du terme d'erreur** ainsi que le meilleur **noyau** à utiliser.

C correspond à la **constante de *soft margin***. Cet hyperparamètre influe sur le **compromis entre l'erreur en apprentissage et la maximisation de la marge**. Plus sa valeur est petite, plus le modèle sera tolérant envers les exemples mal classés ou dans la marge, ce qui correspond à élargir cette dernière. Nous faisons varier C parmi **100 valeurs entre 0.1 et 10** équitablement réparties sur une échelle logarithmique.

De plus, nous cherchons le noyau le plus adapté : ***Radial Basis Function* (rbf) ou un noyau polynomial (poly)**. Cet hyperparamètre définit le **type d'hyperplan** utilisé pour séparer les données.

La figure 1 contient les **hyperpamètres optimisant le SVM pour chaque métrique, avec ou sans PCA**. Nous constatons que le **noyau rbf** permet d'obtenir les meilleures performances dans chaque cas, et que les **valeurs de C maximisant les scores en AMS sont similaires à celles maximisant l'accuracy**.

	Accuracy		AMS	
	sans PCA	avec PCA	sans PCA	avec PCA
C	0.98	0.74	0.89	1.71
Noyau	rbf	rbf	rbf	rbf

TABLEAU 1 – Paramètres optimaux pour le SVM

Pour les **méthodes d'ensemble**, nous faisons varier le **nombre d'estimateurs** (50, 500, 1000 ou 2000), qui correspond à **l'hyperparamètre influant le plus sur la précision de la prédiction**. De plus, nous cherchons la **profondeur maximale optimale des Random Forests**.

Les hyperparamètres optimaux obtenus pour chaque méthode d'ensemble, chaque métrique et en considérant l'intégralité des variables figurent dans le tableau 2. Nous ne remarquons **pas de phénomènes particuliers** : ces hyperparamètres n'évoluent pas de façon monotone lorsque nous passons d'une méthode/métrique à une autre. Dans le cas des données réduites à 20 composantes, les hyperparamètres optimaux obtenus sont, pour la plupart, différents des précédents et, une fois de plus, nous ne constatons pas d'évolution particulière. Nous décidons donc de ne pas les afficher.

	Bagging		RF		RFNan		AdaBoost	
	Accuracy	AMS	Accuracy	AMS	Accuracy	AMS	Accuracy	AMS
Nombre d'estimateurs	500	500	500	500	1000	2000	1000	50
Profondeur max			∞	20	50	∞		

TABLEAU 2 – Paramètres optimaux pour les méthodes d'ensemble (sans PCA)

Sur la figure 5 est tracé **l'AMS moyen obtenu en apprentissage en fonction de la valeur de C (sans PCA)**. Nous remarquons que le score augmente jusqu'à $C \approx 4$ puis stagne.

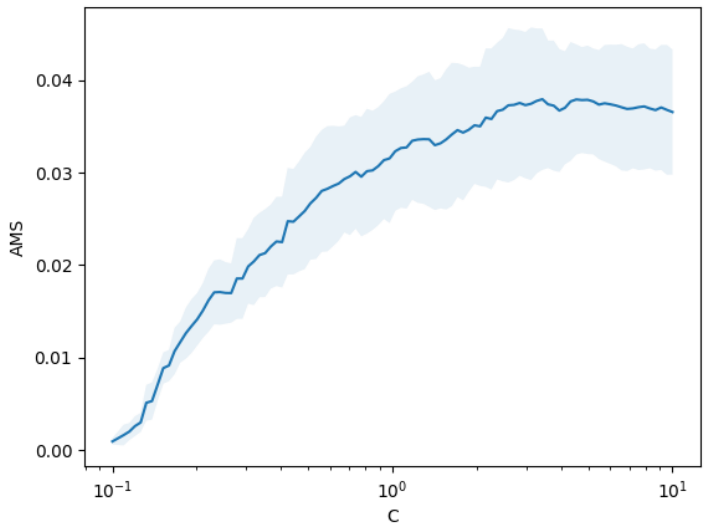


FIGURE 5 – AMS moyen obtenu avec le SVM en fonction de la valeur de C

Les tableaux 3 et 4 contiennent les **scores en accuracy et AMS obtenus avec chaque méthode et, respectivement, sans et avec PCA**. Pour chaque métrique et chaque ensemble sur lequel la performance est calculée, le meilleur score est surligné en vert.

	% Accuracy		AMS (10^{-2})	
	Apprentissage	Test	Apprentissage	Test
SVM	75.90 ± 1.67	75.00 ± 1.16	4.35 ± 0.39	3.19 ± 0.52
Bagging	71.90 ± 1.29	72.60 ± 0.98	3.16 ± 0.60	2.23 ± 0.24
RF	80.70 ± 1.24	78.70 ± 1.59	5.07 ± 0.96	3.98 ± 0.16
RFNan	80.20 ± 0.44	81.30 ± 0.62	5.15 ± 1.12	4.00 ± 0.15
AdaBoost	73.30 ± 0.76	73.30 ± 1.49	3.92 ± 0.55	3.42 ± 0.16

TABLEAU 3 – Performances obtenues avec chaque méthode (sans PCA)

	% Accuracy		AMS (10^{-2})	
	Apprentissage	Test	Apprentissage	Test
SVM	76.30 ± 1.49	75.60 ± 0.45	4.38 ± 0.36	3.02 ± 0.12
Bagging	71.70 ± 1.93	70.30 ± 0.95	2.20 ± 0.49	1.91 ± 0.06
RF	77.20 ± 1.48	74.60 ± 1.37	3.78 ± 0.56	2.69 ± 0.17
AdaBoost	68.00 ± 1.30	68.00 ± 0.99	3.04 ± 0.26	2.05 ± 0.26

TABLEAU 4 – Performances obtenues avec chaque méthode (avec PCA)

Les performances obtenues avec PCA sont toujours inférieures à celles obtenues sans. **L'utilisation d'une PCA n'est donc pas judicieuse**.

Les meilleurs modèles sont RF et RFnan, leur différence de score n'est pas suffisante pour pouvoir les départager. En relançant les expériences sur 10 fois plus de données

(le temps d'exécution devient trop important au delà), nous obtenons les performances AMS (multipliées par 10^{-2}) suivantes :

— RF : 18.20 ± 1.39

— RFnan : 18.29 ± 1.75

Les scores obtenus sont encore une fois très similaires, **nous préférerons cependant le modèle RFnan qui ne nécessite pas d'entraîner des régresseurs pour compléter les données lors du prétraitement**.

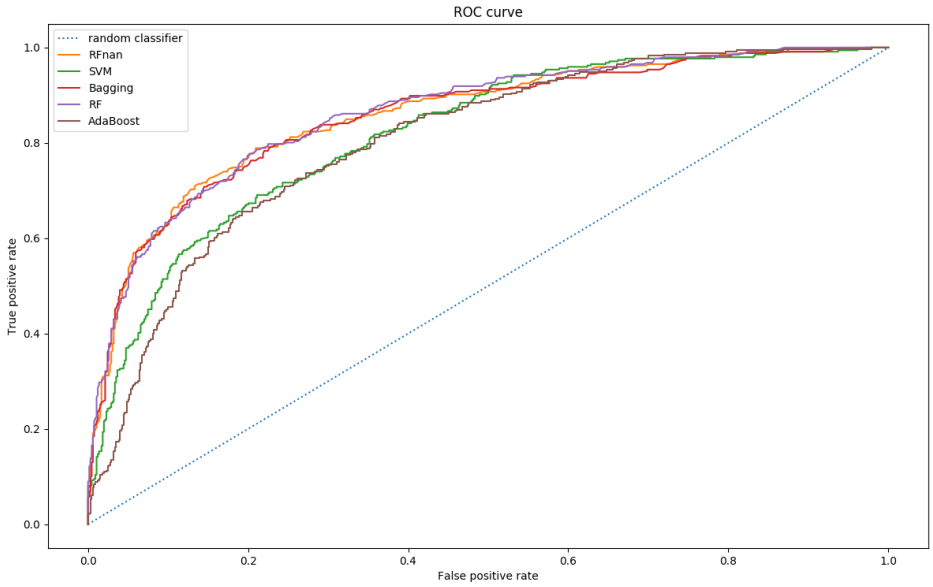


FIGURE 6 – Courbes ROC pour chaque méthode

La figure 6 contient les **courbes ROC** (i.e. le taux de vrais positifs en fonction du taux de faux positifs) **obtenues avec chaque méthode** ainsi qu'avec un classifieur aléatoire. Nous constatons que les courbes correspondant au **SVM et à AdaBoost** sont **plus éloignées du coude du classifieur idéal** (qui passe de $(0, 0)$ à $(0, 1)$ à $(1, 1)$) que les autres.

Afin d'avoir une mesure de qualité d'un modèle indépendante des seuils de classification, nous calculons également **l'aire sous la courbe ROC (AUC) correspondant à chaque méthode** (tableau 5). Les modèles **RF et RFNan obtiennent les meilleurs scores** : ce sont ceux qui distinguent le mieux les deux classes, i.e. ils sont meilleurs que les autres pour étiqueter le background comme étant du background, et le signal comme étant du signal.

	% AUC
SVM	81.95
Bagging	85.5
RF	86.20
RFNan	85.88
AdaBoost	80.45

TABLEAU 5 – AUC obtenus avec chaque méthode

4 Conclusion