UNIVERSITE PARIS 1 PANTHEON-SORBONNE
2020-2021

LES DETERMINANTS D'UN HOBBY SPORTIF CHEZ LES ENFANTS :

RAPPORT DE PROJET

Projet réalisé par :
Djama MAHDI DJAMA

Table des matières

1 Introduction

En France, selon l'INSEE, près d'un tiers des hommes et des femmes tous âges confondus disent pratiquer au moins une activité sportive par semaine. Aujourd'hui avec la fermeture des salles de sports, clubs et associations due à la covid, ce nombre a fortement baissé.

Pour les populations plus jeunes (entre 12 et 17 ans), la pratique d'un sport est relativement plus présente. Environ 77% des garçons et 60% des filles disent pratiquer une activité sportive en dehors des cours d'EPS dispensés dans leurs écoles. Au point même d'être le loisir le plus pratiqué des jeunes Français.

Pour ce qui est des populations très jeunes (entre 6 et 14 ans), les statistiques concernant la pratique sportive ou les autres loisirs sont très peu nombreuses. Ces populations sont souvent oubliées des enquêtes statistiques.Nous axerons donc notre mémoire sur cette population d'âge.

L'objectif de celui-ci sera d'identifier dans un premier temps la proportion de jeunes ayant pour hobby principal la pratique d'un sport. Et de connaître si possible, les éventuelles causes de ce choix (parents sportifs, milieu qui favorise la pratique d'un sport, le sexe ...). A terme, nous essaierons de prédire de la manière la plus efficace possible, le fait d'avoir comme hobby principal la pratique d'un sport chez les jeunes de 6 à 14 ans.

Pour réaliser ceci, nous utiliserons la base de données provenant de l'enquête « les loisirs culturels des 6 à 14 ans » réalisée en France entre 2001 et 2002. Ainsi, dans un premier temps, nous commencerons par présenter rapidement l'enquête, l'entité qui l'a faite, le fonctionnement de l'enquête, ainsi que les variables de la base de données.

Ensuite, nous étudierons la base de données afin de rechercher d'éventuelles données manquantes ou aberrantes, dans le but d'apporter les corrections nécessaires pour pouvoir exploiter correctement celle-ci.

De plus, avant de commencer la modélisation, nous analyserons les données dont nous disposons afin d'avoir une première idée de l'influence que pourrait avoir chacune des variables, la probabilité d'avoir un hobby de sport chez les jeunes.

Enfin nous proposerons une comparaison de plusieurs modèles prédictifs permettant de prédire notre variable à expliquer.

2 L'enquête

L'enquête « loisirs culturels des 6 à 14 ans » offre une vision panoramique des loisirs, et pratiques culturels choisies par les enfants de 6 à 14 ans (Jeux vidéo, télévision, musique, sport…). En plus de proposer un vaste nombre d'informations concernant la famille, le milieu d'origine et l'école des enfants présent dans l'enquête.

Celle-ci a été réalisée par le DEPS (Département des études de la prospective et des statistiques) auprès de 3306 familles entre 2001 et 2002.Pour la réalisation de l'enquête, le DEPS a tiré au hasard 180 écoles et collèges, stratifié suivant leur démographie (partition publique/privée, ZEP, tranche d'unité urbaine, taille…).

Le questionnaire contient à lui seul plus de 400 questions, réparties en plusieurs catégories. On retrouve par exemple des questions sur les caractéristiques des enfants (sexe, tranche d'âge, classe…), les caractéristiques des parents (CSP, tranche d'âge, situation maritale…), ainsi qu'une multitude de questions sur les loisirs des familles (loisirs des enfants, des parents, si les parents pratiquent avec les enfants…).

3 La base de données

La base de données de l'enquête contient 1582 variables et 2904 observations. Le nombre de variables est beaucoup plus élevé que le nombre de questions, car certaines questions ont été divisées en plusieurs variables.

Les variables correspondent aux réponses des familles données à l'enquête. Catégorisées de la même façon que les questions.

Toutes les variables sont catégoriques, mais sont affichées sous format « numérique », c'est-à-dire qu'elles sont divisées en plusieurs modalités, prenant 1 si l'évènement 1 se réalise, 2 si l'évènement 2 etc.
Il n'y a donc aucune variable de type numérique à proprement parler.

4 Traitement sur la base

4.1 Données manquantes

Au vu du grand nombre de questions présent dans l'enquête, une part non négligeable de familles n'a pas complété le questionnaire.

Ainsi, sur les 1582 variables, la moyenne de données manquantes par variable est de 52%. Plus de la moitié des données par variables sont manquantes. Seules 18 variables n'ont présente aucune. Notre variable cible (le hobby des enfants) possède quant à elle 54% de données manquantes.

Comme nous comptant effectuer un certain nombre d'algorithmes sur cette base de données, la forte proportion de données manquantes pose un gros problème. La plupart des algorithmes (voir pratiquement tous) ne se lancent pas avec des données manquantes.

Pour pallier à ça nous commençant d'abord par supprimer les variables possédant plus de 60% de données manquantes. La décision de ce seuil a été faite en fonction du pourcentage de données manquantes de notre variable cible (54%). Après cette opération, le nombre de variables restant est descendu à 1062.

Ensuite, comme nous ne voulions pas supprimer d'autres variables, nous avons décidé d'opter pour un remplacement des données manquantes par algorithme, car les variables de la base sont catégoriques, les remplacer par leurs moyennes ou leurs médianes est donc de ce fait totalement absurde.

L'algorithme que nous choisissons est celui du KNN. L'algorithme du KNN (K plus proches voisins) fait partie de l'apprentissage supervisé. Il est utile pour les problèmes de classification, mais peut aussi être utilisé comme algorithme de régression. Cette méthode a pour but de classifier des points cibles (classe méconnue) en fonction de leurs distances par rapport à des points constituant un échantillon d'apprentissage (c'est-à-dire dont la classe est connue a priori).

Nous l'utiliserons ici, car celui-ci est efficace sur les variables catégoriques. En ce qui concerne le nombre de voisins que nous avons choisis pour le modèle, nous le mettons à 5 pour commencer. Pour savoir si le modèle a performé relativement bien, nous décidons de comparer l'effectif de modalités de notre variable cible (le hobby des enfants) sans ajout de données par algorithme et avec.

	Avec Algorithme		Sans Algorithme
17.0	338	17.0	206
34.0	338	34.0	172
16.0	196	16.0	144
24.0	181	9.0	131
22.0	178	33.0	118
39.0	166	46.0	76
41.0	160	44.0	59
9.0	152	10.0	44
5.0	151	32.0	33
10.0	139	42.0	32
42.0	122	18.0	27
33.0	120	39.0	23
44.0	89	41.0	23
7.0	86	2.0	22
21.0	78	24.0	21
46.0	76	38.0	21
2.0	60	28.0	19
32.0	54	47.0	13
18.0	27	21.0	12
38.0	25	5.0	11
3.0	21	43.0	7
28.0	21	36.0	7
4.0	17	3.0	7
47.0	13	58.0	6
58.0	10	26.0	6
12.0	10	7.0	5
20.0	7	12.0	5
36.0	7	45.0	5
43.0	7	20.0	5
54.0	6	54.0	5
26.0	6	25.0	5
35.0	5	35.0	4
45.0	5	6.0	4
25.0	5	4.0	4
57.0	4	57.0	4
6.0	4	22.0	4
15.0	4	15.0	4
52.0	3	52.0	3
14.0	2	30.0	2
30.0	2	8.0	2
8.0	2	14.0	2
13.0	1	29.0	1
29.0	1	13.0	1
19.0	1	23.0	1
27.0	1	50.0	1
50.0	1	27.0	1
23.0	1	19.0	1
53.0	1	53.0	1

Figure 1 : Comparaison entre algorithme d'imputation de données manquante et sans

Ainsi, nous remarquons que la répartition de données manquantes par modalité sans algorithme correspond plus ou moins à celle avec algorithme. L'algorithme du KNN avec 5 voisins a donc fourni un résultat convenable, nous décidons de garder cette base pour la suite de l'étude. Enfin, nous n'avons remarqué aucune donnée aberrante dans la base.

4.2 Traitement sur la variable cible

La variable cible que nous choisissons est la P21, celle-ci contient le hobby des enfants réparti en 59 modalités allant de la télévision au VTT. Afin de répondre à notre problématique qui consiste à la prédiction d'un hobby sportif chez les enfants, nous décidons de transformer cette variable en une variable dichotomique univariée.

Ainsi, nous considérons comme un sport, toutes les modalités comprises dans la liste ci-dessus :

- Foot
- Marche/Randonnée
- Sport
- Sport collectif
- Sport individuel
- Natation/Piscine
- Sport mécanique
- Vélos/VTT
- Équitation
- Sport nautique
- Sport de combat
- Sport de glisse

Sur 2904 observations, 752 possèdent un hobby parmi la liste ci-dessus, soit environ 26% de la base.

5 Statistiques descriptives

5.1 Test statistique à grande échelle

Dans cette section, nous étudierons la relation entre notre variable expliquée et nos variables explicatives.

Comme nous avons beaucoup de variables et afin de vérifier s'il existe des variables qui ont un effet sur notre variable expliquée, nous décidons de réaliser un test de Khi2 entre la variable hobby sportif et l'ensemble de nos variables explicatives. Pour rappel, les hypothèses du test de khi2 sont les suivantes :

Hypothèses du test :

Ho = la distribution observée n'est pas significativement différente de la distribution théorique.

H1 = la distribution observée est significativement différente de la distribution théorique.

Nous rejetons H1 si la p-value est inférieure à 5%, dans notre cas, 588 variables n'ont pas d'effet sur notre variable expliquée. Nous décidons de nous séparer de celles-ci pour réduire le pool de variables déjà conséquent, à ce stade du projet, il nous reste 474 variables.

5.2 Visualisation de certaines variables

Dans cette section, nous visualiserons par le biais de graphiques, la relation entre certaines variables explicatives et notre variable à expliquer.

Sur les 474 variables, nous en avons sélectionné 3 pour rapidement voir l'effet du hobby sportif sur ces variables.

Le sexe de l'enfant :

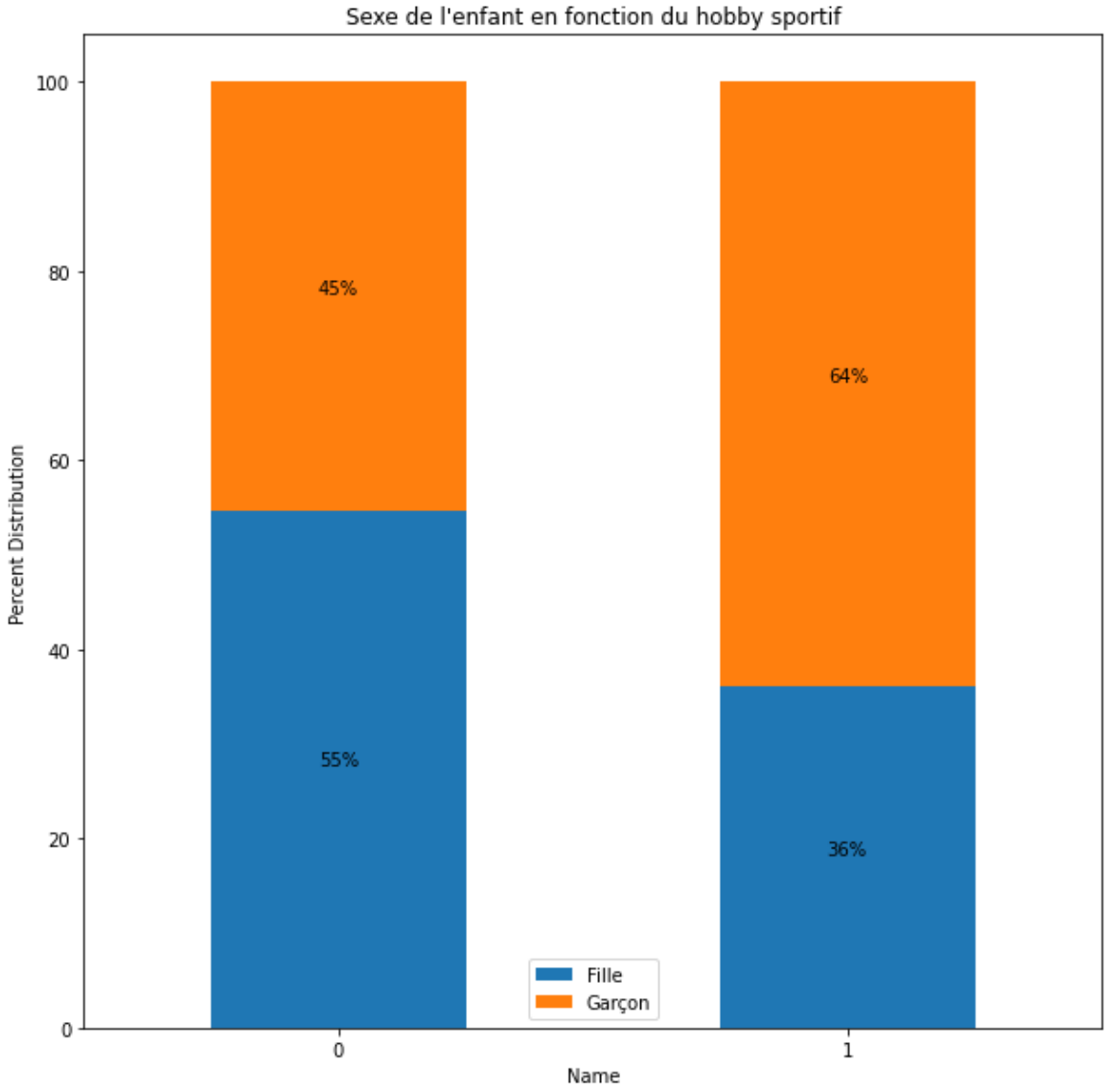


Figure 2 : Sexe de l'enfant en fonction du hobby sportif

Le graphique ci-dessus nous présente la proportion de filles/garçons ayant ou n'ayant pas un hobby sportif dans la base. Celui-ci nous démontre que les filles sont celles avec le moins de hobbys sportifs. Ce constat concorde avec les statistiques nationales de l'INSEE.

Le fait d'avoir une famille biparentale et monoparentale :

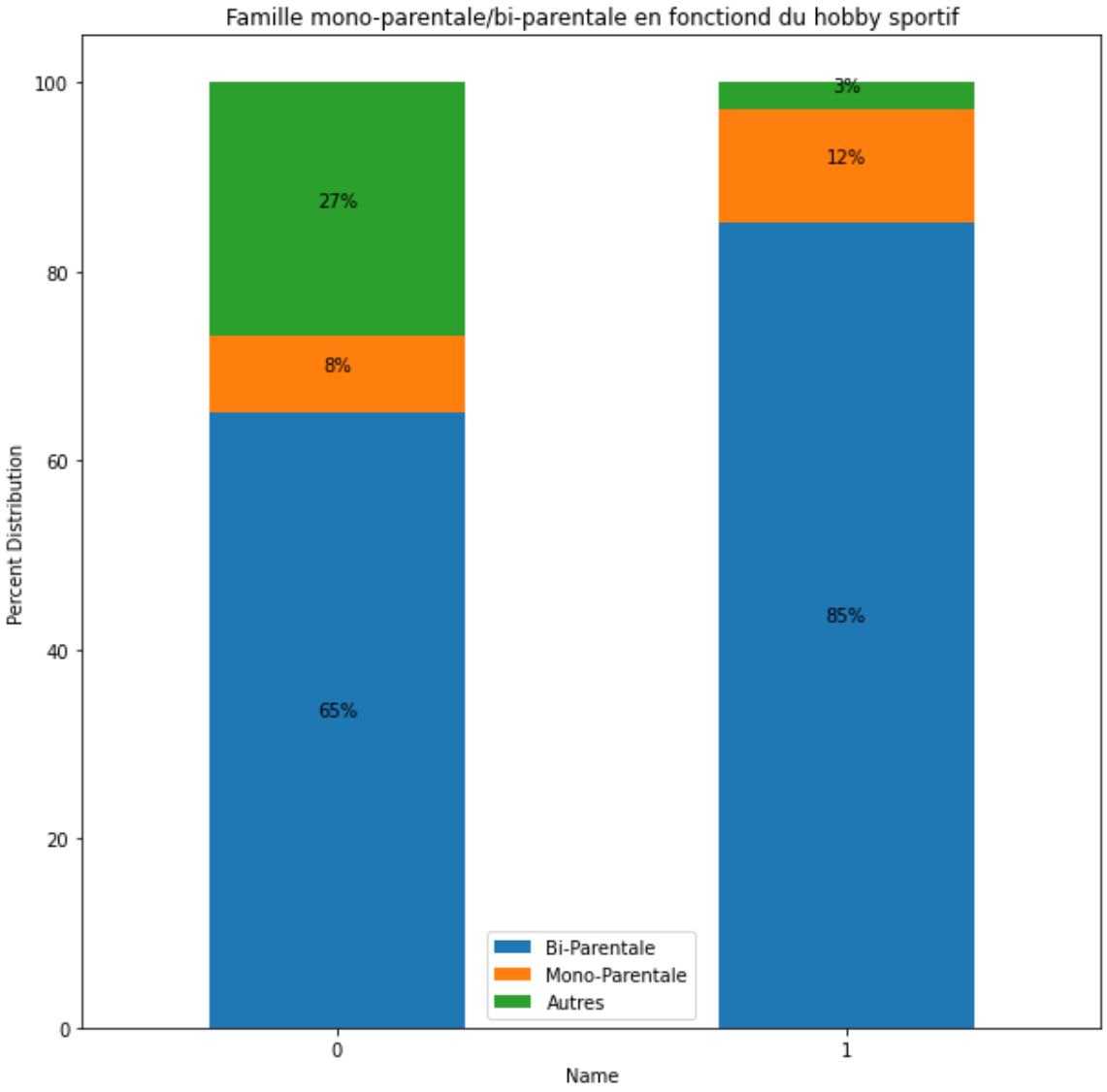


Figure 3 : La fait d'avoir une famille biparentale/monoparentale en fonction du hobby sportif chez les enfants

Ainsi, d'après le graphique ci-dessus, les familles biparentales sont celles possédant le plus d'enfants avec un hobby sportif. Nous pouvons supposer que comme l'enfant possède 2 parents, il y a plus de chance que l'un des deux soit sportif et le pousse à en faire aussi.

Le fait d'avoir un parent sportif :

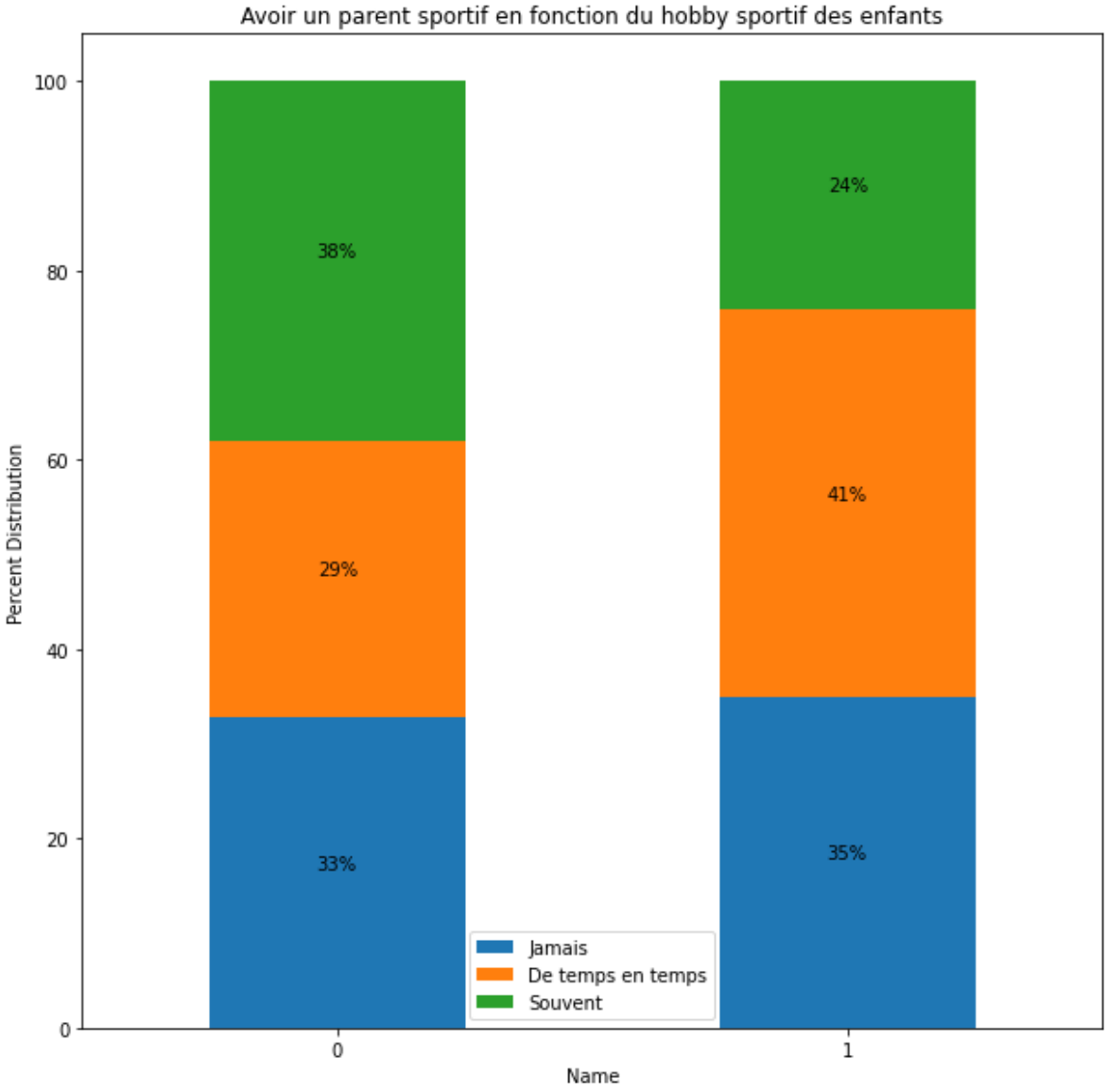


Figure 4 : Le fait d'avoir un parent sportif en fonction du hobby sportif des enfants

Étonnamment, le fait d'avoir un parent sportif ne semble pas influer sur notre variable à expliquer. La proportion de parents qui ne font jamais de sport est la même pour les 2 modalités de notre variable à expliquer.

6 Modélisation

Dans cette section, nous effectuerons et comparerons un certain nombre d'algorithmes afin de prédire notre variable à expliquer avec nos variables explicatives

6.1 Étape préalable

Avant de commencer la modélisation, il nous faut préparer la base de données pour qu'elle soit correctement exploitable.

Dans un premier temps, nous transformant toutes nos variables catégoriques en variables dichotomiques univariées prenant 1 si l'évènement se réalise et 0 sinon pour chacune des modalités. Nous faisant ceci pour pouvoir exploiter correctement chacune des variables catégoriques dans des algorithmes sur python. Le codage en variable dummy est l'une des solutions pour utiliser des variables catégoriques dans un modèle.

De plus, afin d'éviter le problème de multi colinéarité (corrélation entre les variables), nous enlevant pour toutes les variables 1 modalité. Le nombre de variables après toutes nos manipulations est de 2212. Comme ce nombre est extrêmement élevé, nous décidons de le tronquer en effectuant un Random Forest (algorithme de l'apprentissage supervisé que nous verrons plus bas), celui-ci nous a permis de sélectionner les variables ayant un taux d'importance au-dessus de 0.002 (donc non égale à 0), ce qui nous évitera de faire tourner le modèle avec des variables qui n'ont aucune importance. Cette méthode nous a permis de passer de 2212 variables à seulement 34.

Enfin, concernant la séparation du jeu de données en train/test, nous décidons de partir sur une proportion 30% pour les données de test, car notre base possède relativement peu de données. De plus, comme la variable à expliquer est déséquilibrée (25% d'enfants avec un hobby sportif), nous introduisons la notion de séparation stratifiée, qui consiste à séparer le jeu de données avec la même proportion de hobbys sportifs dans la base de test et de train. Ceci nous permettra d'éviter les situations de sur-apprentissage.

6.2 Algorithmes :

Arbre à décision :

L'algorithme d'arbre de décision (decision tree) fait partie de l'apprentissage supervisé. Celui-ci est utilisable à la fois pour la classification et la régression, mais nous verrons ici que sa partie classification. L'algorithme décompose un ensemble de données en sous-ensembles de plus en plus petits, tandis qu'en même temps, un arbre de décision associé est progressivement développé. Il est cependant réputé pour être assez simpliste et a tendance à surajustement, car tout est imbriqué en un seul arbre.

Concernant le choix des paramétrages optimaux, nous décidons de partir sur 3 approches différentes pour commencer. Une première méthode aléatoire qui consiste à faire varier de manière aléatoire un certain nombre de paramètres, puis sélectionner le meilleur ensemble de paramètres en prenant le AUC (C'est une mesure de la probabilité pour que le modèle classe un exemple positif aléatoire au-dessus d'un exemple négatif aléatoire) max.

La seconde méthode est un grid search. Le grid search va croiser l'ensemble des paramètres et construire un modèle pour chacun de ceux-ci, de ce fait cette méthode est assez couteuse en temps. Les meilleurs paramètres seront sélectionnés avec le score grid search.

Enfin, la dernière méthode consiste à effectuer un bayes search. Cette méthode consiste à construire un modèle probabiliste de la fonction qui relie les valeurs des hyperparamètres à l'objectif évalué sur un ensemble de validation en évaluant itérativement une configuration de paramètres prometteurs sur la base du modèle actuel, le but sera de chercher l'optimum. Les meilleurs paramètres seront sélectionnés avec le score d'optimisation.

Afin de comparer tous les modèles obtenus à l'aide des 3 méthodes, nous utiliserons la courbe ROC avec l'AUC et la courbe précision/rappel. La courbe ROC est un graphique représentant les performances d'un modèle de classification pour tous les seuils de classification. Cette courbe trace le taux de vrais positifs en fonction du taux de faux positifs.

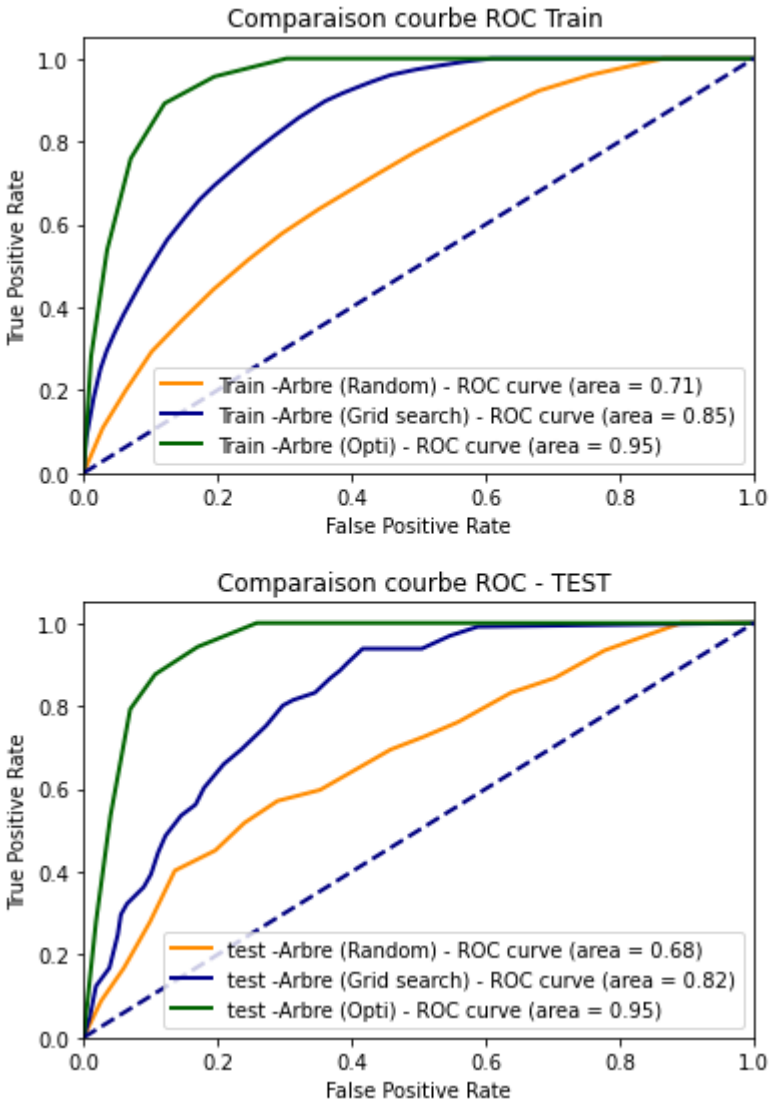


Figure 5 : Comparaison courbe ROC Train/Test

Ci-dessus se trouve la comparaison entre la courbe ROC du jeu de données d'entraînement et celle sur le jeu de données de test.

D'après les deux graphiques, la méthode ayant trouvé le modèle le plus performant est celle du bayes search. Le taux de vrais positifs en fonction du taux de faux positifs y est le plus élevé (l'AUC est de. De plus, nous remarquons que les modèles sont relativement stables étant donné que les courbes sont plus ou moins de la même forme.

Comme le modèle avec bayes search a bien performé, nous utiliserons cette méthode comme référence pour les prochains algorithmes que nous utiliserons. La méthode aléatoire procure de moins bons résultats et le grid search est très contraignant au niveau de son temps d'exécution.

Pour aller plus vite, nous comparerons maintenant l'ensemble des 7 modèles que nous avons effectué dans les mêmes graphiques pour sélectionner le modèle le plus performant.

Les 7 modèles sont les suivants :

- L'arbre à décision que nous avons étudié plus haut
- La régression logistique (un algorithme qui fait partie de l'apprentissage supervisé. Il est utilisé en majorité pour des problèmes de classification binaire)
- Le random forest (cet algorithme fait lui aussi partie de l'apprentissage supervisé. Tout comme le decision tree, il est utilisable pour les problèmes de classification et de régression. Contrairement au decision tree, l'algorithme est plus complexe, il réduit le surajustement causé par un seul arabre
- . L'algorithme permet aussi de calculer l'importance des variables et est meilleur pour les problèmes de prédictions. Nous nous attendons donc à de meilleurs résultats que l'algorithme de decision tree).
- Le Bagging (C'est un algorithme d'apprentissage d'ensemble qui combine les prédictions de plusieurs arbres de décision
- Le Xgboost (C'est une implémentation open source optimisée d'arbres de boosting de gradient)
- Adaboost (C'est un algorithme de Boosting qui combine plusieurs classifieurs peu performant pour amplifier leurs résultats)
- Stacking (C'est un algorithme qui permet de combiner plusieurs modèles de classification via un meta-classifieur). Celui-ci va nous permettre de combiner les 6 algorithmes vus précédemment.

Nous représenterons la courbe ROC pour chacun des modèles ci-dessus.

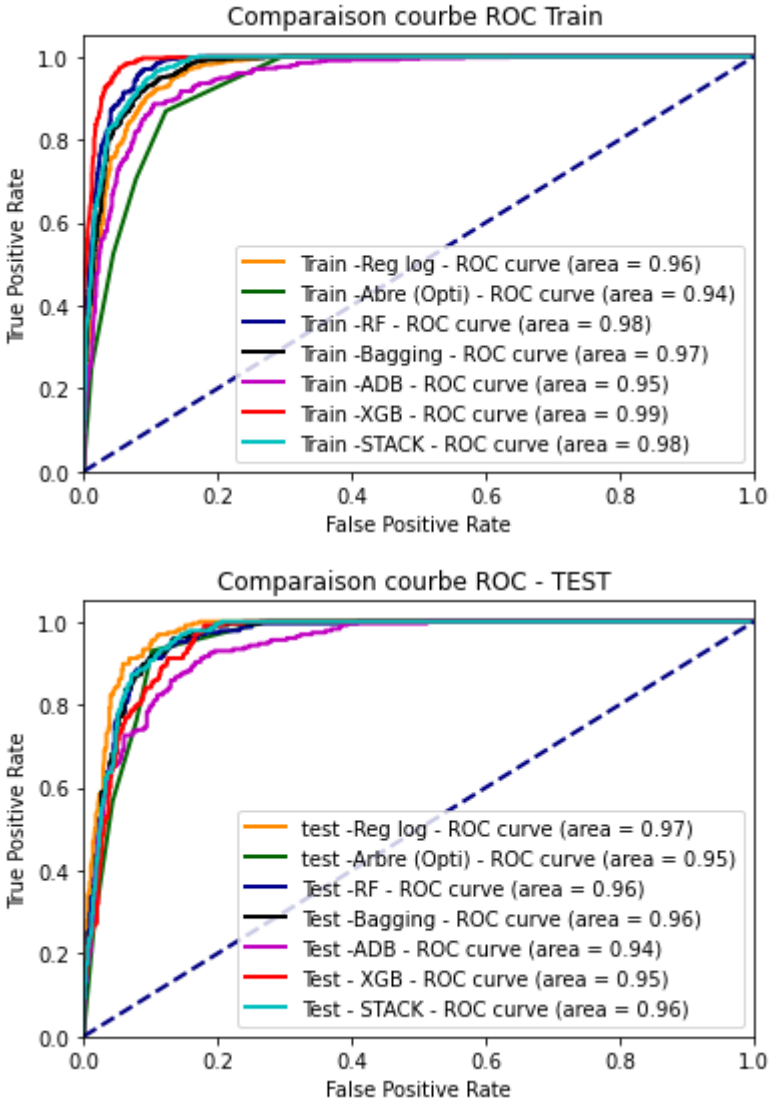


Figure 6 : Comparaison courbe ROC Train/Test

D'après les graphiques ci-dessus, tous les modèles sont globalement très performants, l'AUC est supérieur à 0,90 pour l'ensemble des algorithmes. De plus, ceux-ci sont relativement stables, les courbes ROC Train/Test sont comparables et de la même courbure. Les modèles sont donc globalement satisfaisants.

Celui qui semble se distinguer des autres est le Stacking, son AUC est de 0.96. Ceci semble cohérent, car il s'agit de la combinaison des 6 autres algorithmes que nous avons choisis. C'est le modèle que nous choisirons.

6.3 Performance du modèle choisi

Dans cette section, nous étudierons les performances du stacking que nous avons effectué.

Courbe précision/rappel :

Avant toute chose, nous allons établir un seuil de cut off. Le cut off est la valeur seuil telle que si Score >=cut off alors Ypred=1. Par défaut, dans sckit-learn le cut off sera 0.5. Pour établir celui-ci, nous étudierons la courbe précision/rappel du modèle.

Pour évaluer les performances d'un modèle de façon complète, il est nécessaire d'analyser **à la fois** la précision et le rappel. Malheureusement, précision et rappel sont fréquemment en tension. Ceci est dû au fait que l'amélioration de la précision se fait généralement au détriment du rappel et réciproquement.

La précision mesure le pourcentage d'**enfants avec un hobby sportif** ayant été classifié correctement, le rappel mesure le pourcentage d'enfants sans hobby sportif ayant été classifiés correctement.

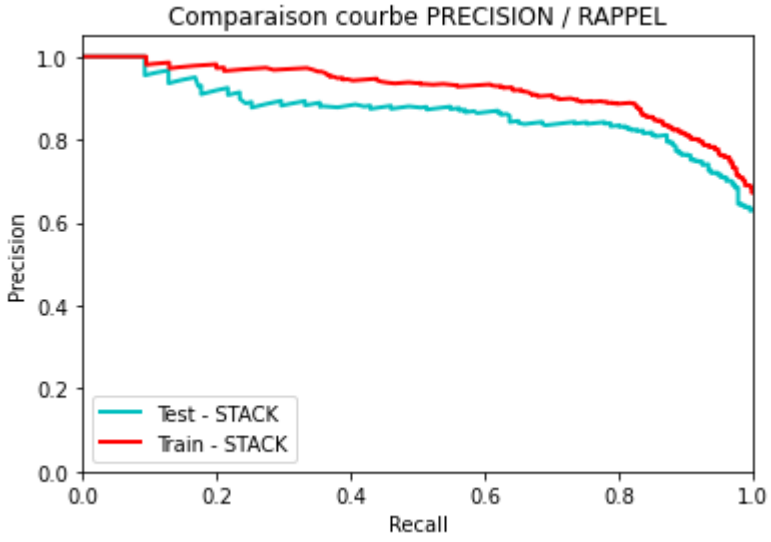


Figure 7 : Comparaison courbe Précision/Rappel

Le graphique ci-dessus présente la courbe de précision en fonction du rappel. Nous n'utiliserons pas la courbe train pour évoluer le modèle, car celle-ci est en situation de surapprentissage. En regardant la courbe train, et pour une précision de 95% nous prendrons un rappel de 20%. C'est-à-dire que nous allons cibler 20% de la population des enfants avec un hobby sportif pour 95% de précision.

Matrice de confusion :

Pour évaluer la performance du modèle, nous effectuerons une matrice de confusion avec les prédictions du jeu de données de test.

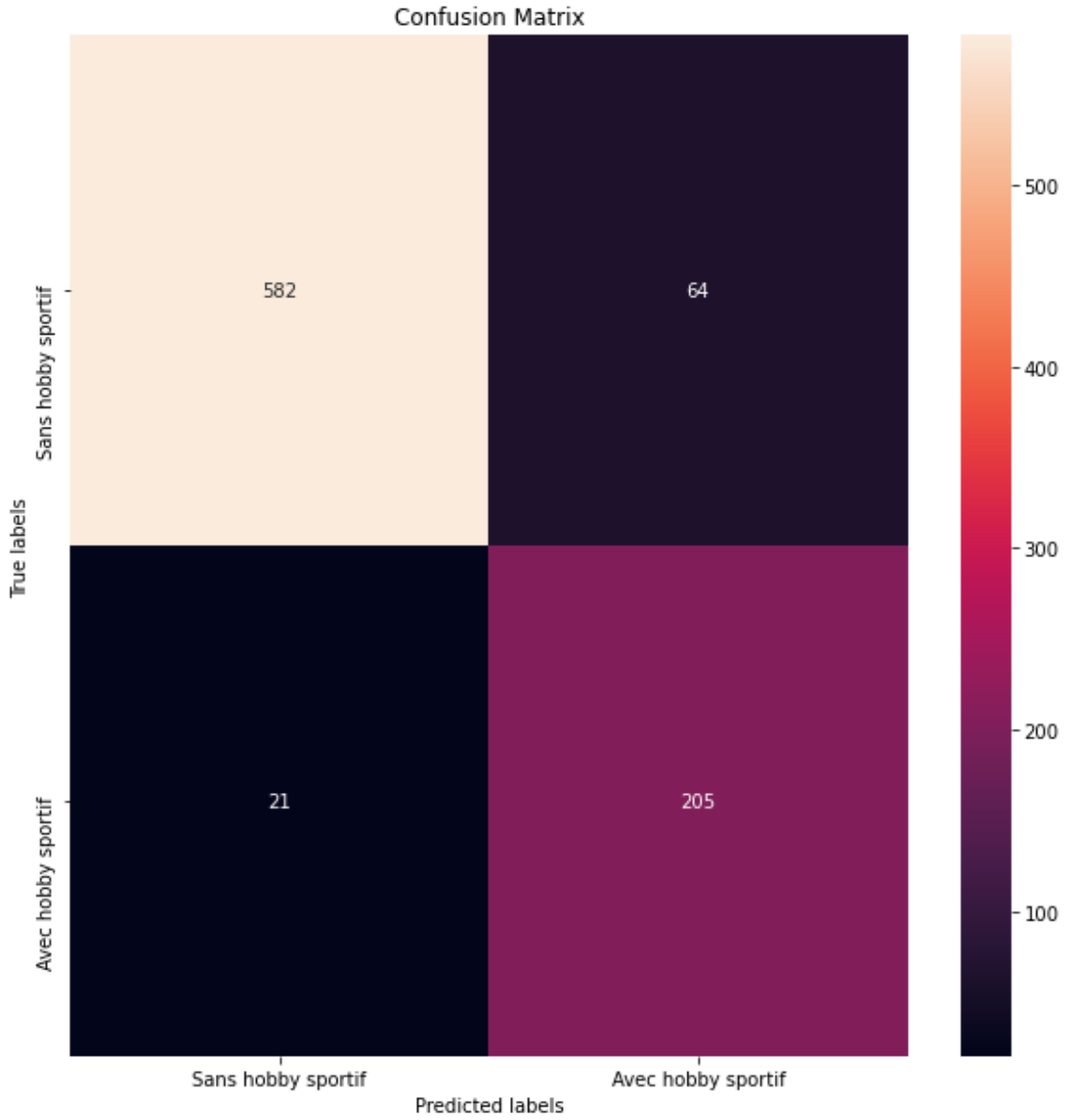
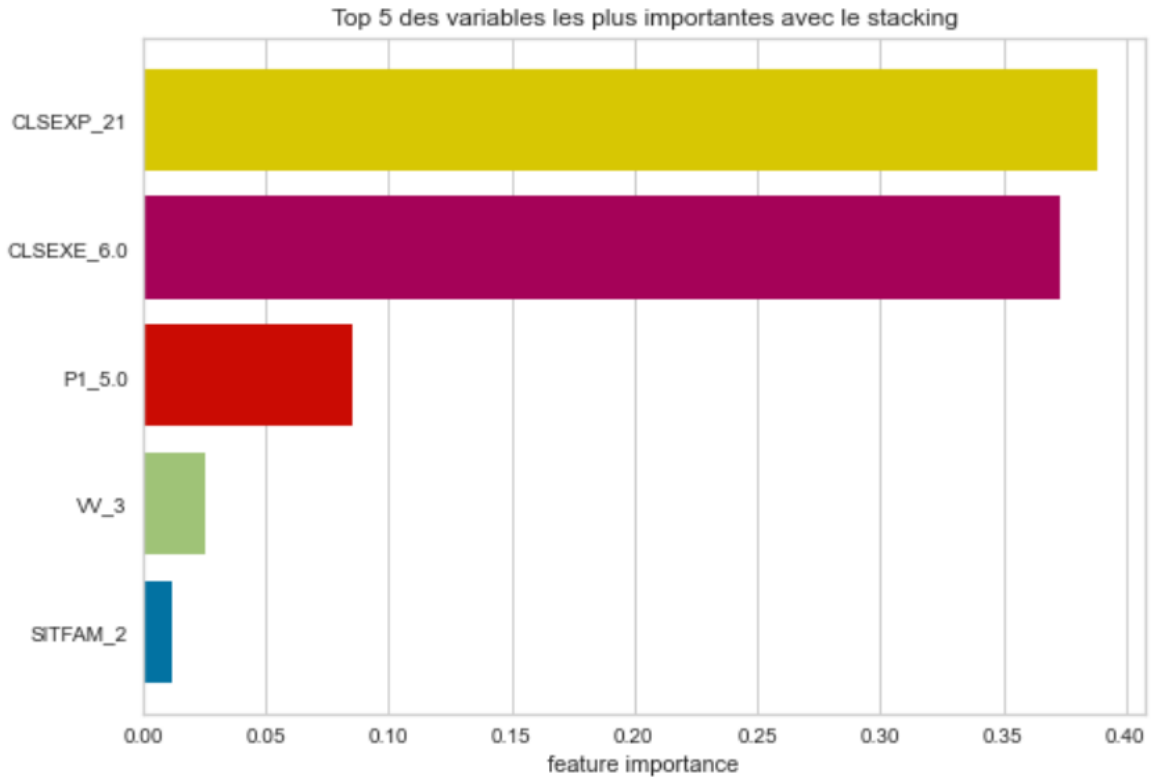


Figure 8 : Matrice de confusion du modèle

Le modèle est globalement très performant, sur les 582 enfants sans hobby sportif de la base de test, celui-ci en a classé 582 dans la bonne catégorie et 21 dans la mauvaise. Soit un taux d'erreur (précision) d'environ 4%. Pour ce qui est de la classification des enfants avec hobby sportif, celui-ci a classifié correctement 205 de ceux-ci, et incorrectement 64. Le taux d'erreur (rappel) est d'environ 22%. Ce qui correspond presque parfaitement à ce que nous avions déterminé plus haut avec la courbe précision/rappel.

Variables les plus importantes :



Nous représentons ci-dessus les 5 variables qui agissent le plus sur le modèle (donc les 5 déterminants les plus importants dans le fait d'avoir un hobby sportif ou non).

La première variable correspond à la modalité 21 de la variable CLSEXP (Classe et cursus de l'enfant par rapport au sexe), qui correspond au total d'enfants au Collège (Garçon+Fille). Donc le fait d'être collégien est le plus grand déterminant sur le fait d'avoir un hobby sportif ou non.

La seconde variable correspond à la modalité 6 de la variable CLSEXE (Classe de l'enfant par rapport au sexe), qui correspond au fait d'être un garçon en 6e.

La 3e P1 (Nb de pièces du logement) correspond à la modalité 5, qui est de posséder 5 pièces dans le foyer.

La 4e VV (Nombre de personnes dans le foyer) correspond à la modalité 3, c'est-à-dire 3 personnes dans le foyer.

La 5e SITFAM_2 (Situation familiale) correspond à la modalité 2, c'est-à-dire vivre avec le père et la mère.

Avec toutes ces informations, nous pouvons en conclure que l'individu type ayant le plus de chances d'avoir un hobby sportif est un enfant en 6e vivant avec ses deux parents dans un logement à 5 pièces.

7 Conclusion

Le but du projet était de prédire de la manière la plus efficace possible le fait d'avoir un hobby sportif chez les enfants de 6 à 14 ans. Après plusieurs manipulations sur la base de données que nous avions, et quelques analyses descriptives, nous avons réussi à créer un modèle prédisant au mieux notre variable à expliquer. Celui-ci est globalement très satisfaisant, à la fois performant et stable, il nous a permis de déceler les principaux déterminants du fait d'avoir un hobby sportif chez les enfants. D'après le modèle, l'individu type ayant le plus de chances d'avoir un hobby sportif est un enfant en 6^{e} vivant avec ses deux parents dans un logement à 5 pièces.