Comparaison d'algorithmes de classification ML pour prédire le cancer.

Abdoulaye Diouma SOW

August 6, 2023

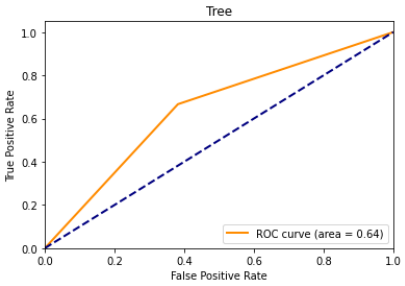
1 Introduction

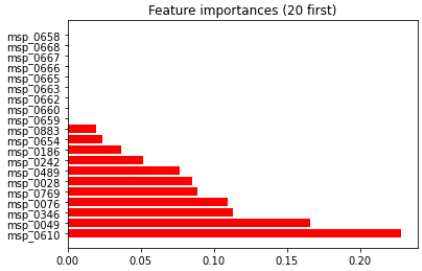
L'objectif du projet est d'effectuer des analyses basées sur l'apprentissage automatique sur les abondances des espèces et de proposer un algorithme de classification ML qui peut prédire le statut des patients à partir de leurs abondances d'espèces. Je propose quatre algorithmes ML différents puis compare les performances des différents modèles et enfin sélectionne le meilleur. Je proposerai une représentation des différentes espèces impliquées dans le modèle retenu sous la forme d'un réseau en utilisant les distances des espèces en fonction de leur rôle comme indiqué dans l'énoncé. La variable cible qui est le statut du patient contient quatre modalité: Cancer, Normal, Small adenoma et Large adenoma. Cette variable d'interet a été convertie en deux categorielles: Cancer et Normal. On suppose que les patients avec Small adenoma ne sont pas dans un stade avancé. Dans l'étude, ces patients ont été considérés comme normal tandisque ceux avec Large adenoma ont été considerés comme Cancéreux. J'ai mis les individus dans le même ordre . L'ordre dans le quel ils etaient dans le fichier meta etait différent de celui de all.sample. C'est pour ne pas biaiser l' étude. L'echantillon 'SAMEA2466920' a été retiré de l'etude car il etait présent dans meta et absent dans all.sample. Pour les mesures de performances, j'ai calculé la precision, le recall, le F1-score et tracé les courbe de roc. Les données ont été separées en deux échantillons: Test et apprentissage. Les modéles sont ajustés sur les données d'apprentissage puis validés sur l'echatillon de test. Les variables d'importance du modéle retenu ont été utilisés pour la representation du reseau. Seules qui apportent des informations ont été selectionnés. J'ai utiliser le logiciel python pour ajuster et valider les modéles puis R pour la classification des especes.

2 Arbre de decision

Les arbres de décision peuvent être utilisés pour des problémes de classification ou de regression. Les arbres de décision sont des modèles non paramétriques :

ils ne sont pas contrôlés par une fonction de décision mathématique et n'ont pas de poids ou d'interception à optimiser. L'analyse de ces deux figures montre

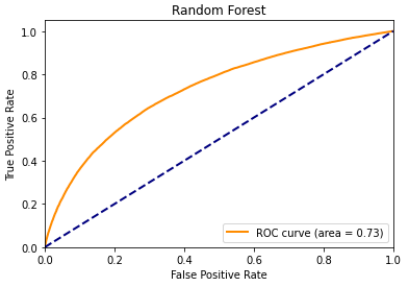


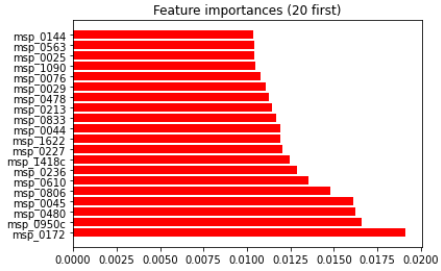


que l'arbre decision présente une faible auc et le nombre de facteurs importants detectes est tres faible.

3 Random Forest

Random Forest est une méthode composés de plusieurs arbres de décision. L'algorithme de Random Forest est constitué d'une collection d'arbres de décision, et chaque arbre de l'ensemble est composé d'un échantillon de données tiré d'un ensemble d'apprentissage avec remplacement, appelé bootstrap

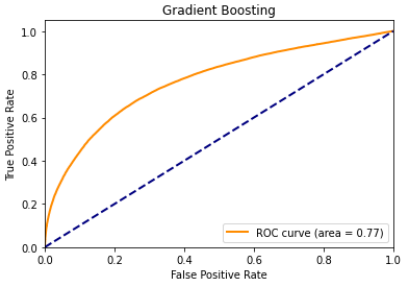


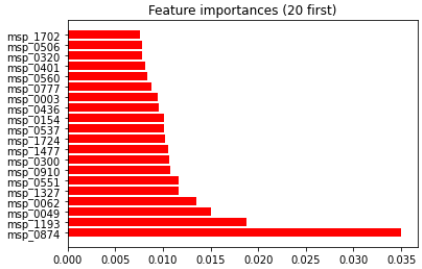


L'analyse des deux figures obetnues à l'aide de Random Forest montre une auc qui est plutot bonne par rapport à celle de l'arbre de decision. Cependant plus de facteur importants sont detectés.

4 Gradient Boost

L'algorithme de Gradient Boosting a beaucoup de points communs avec Adaboost. Tout comme Adaboost, il s'agit d'un ensemble de "weak learners", créés les uns après les autres, formant un strong learner. L'analyse des deux fig-

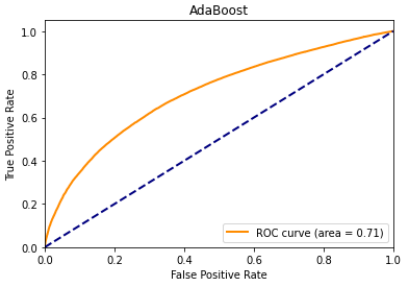


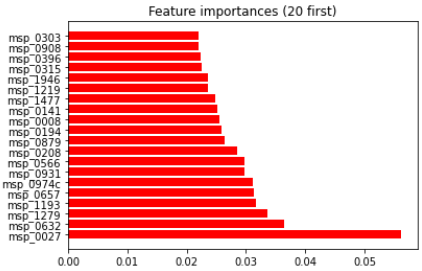


ures montre une meilleure auc rapport à celle de l'arbre de decision et Random Forest. Cependant plus de facteur importants sont detectés.

5 AdaBoost

Les " weak learners " d'AdaBoost sont généralement des arbres décisionnels à seulement deux branches et deux feuilles (aussi appelés souches) mais on peut utiliser d'autres types de classificateur L'analyse des deux figures obetenues à





l'aide d'Ada Boost montre une auc moins bonne par rapport à celle du Gradient Boost. Des facteurs importants sont detectés.

Table 1: Quelque mesures de performance

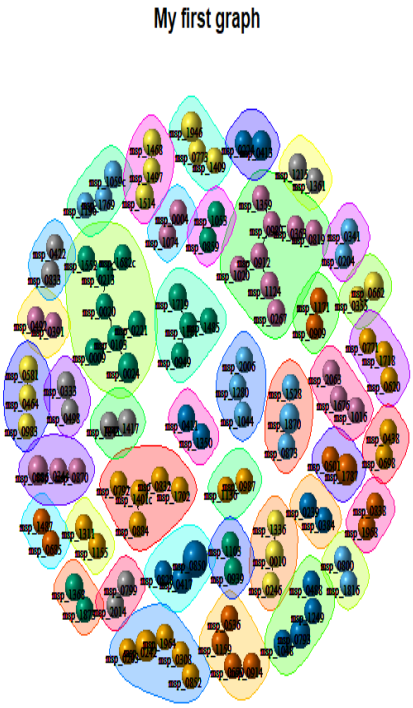
Metrics	Random Forest		Gradient Boost		AdaBoost	
	Cancer	Normal	Cancer	Normal	Cancer	Normal
Precision	0.80	0.68	0.81	0.75	0.71	0.57
Recall	0.74	0.75	0.81	0.75	0.63	0.65
f1-score	0.77	0.71	0.81	0.75	0.67	0.60

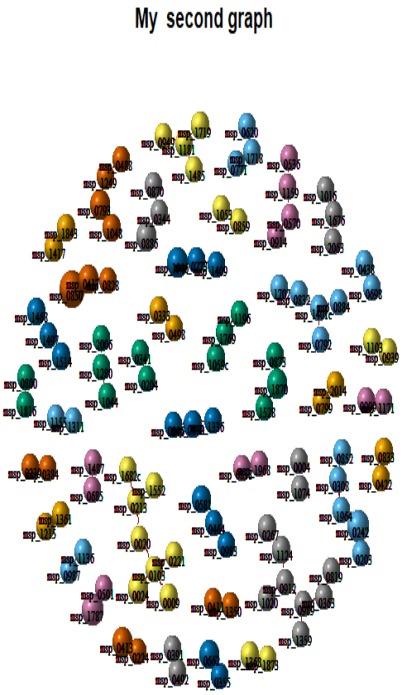
L'analyse des résulats du tableau montre que le Gradient Boost est le meilleur

modéle pour prédire le cancer dans notre cas d'etude. Ce modèle présente la meilleure auc, la meilleure precision, le meilleur recall et le meillleur F1-score.

6 Classification des espéces

Ces resultats ont été obtenus à l'aide package igraph disponible sur R. 747 espéces ont retenues à l'aide du Gradiant Boost comme étant des variables d'importance. Ces espéces ont été filtrées à partir de la matrice d'abondance. On obtient au final une nouvelle matrice d'abondance de 747 lignes et 187 echantillons. La matrice de distance a été construite pour la représentaion du reseau d'interaction espéces éspeces. Des espéces libres c'est à dire non connectées ont été retirées du reseau. Ce qui fait qu'on se retrouve avec centaine d'espéces. L'analyse du reseau montre une bonne classification des espéces. Des





éspeces sont classées en focntion de leurs connexion avec d'autres espéces.