La prédiction des types du cancer par 4 algorithmes de Machine Learning

Auteur : Abdeljalil SENHAJI RACHIK

INTRODUCTION :

Machine Learning (ML) est le domaine de l'informatique, par lequel les systèmes informatiques peuvent comprendre les données de la même manière que les humains. Cependant, machine Learning est une intelligence artificielle qui peut utiliser des algorithmes ou des méthodes pour extraire des modèles à partir de données brutes.
L'objectif principal du ML est de permettre aux systèmes informatiques d'apprendre de l'expérience sans être explicitement programmés ou d'intervention humaine. Avec la progression de séquençage du génome humain, on estime le nombre de gènes codant une protéine chez l'être humain à environ 20 000 et vue l'amélioration des programmes informatiques et le développement de le machine Learning, les chercheurs visent à construire des modèles utilisant l'expression génique (et d'autres caractéristiques) pour prédire les maladies. Ces modèles pourraient avoir deux avantages principaux. Premièrement, il pourrait aider les médecins en installant un pipeline automatique, réduisant ainsi les erreurs de prédiction des maladies à la population générale. Deuxièmement, ces modèles de machine Learning pourraient prédire les maladies avant que les premiers symptômes ne se manifestent, ce qui permettrait d'augmenter les chances de survie.
Dans ce rapport, 4 modèles machine Learning utiliser pour la prédiction du cancer ont été construit en utilisant des gènes qui code pour 20531 des protéines humaines. Plus précisément, on prédit la classe de 5 cancers LUAD, COAD, PRAD, KIRC et BRCA, respectivement le cancer du sein, le cancer du rein et du rein, le cancer colorectal, le cancer du poumon et le cancer de la prostate en fonction du niveau d'expression de ces gènes chez 801 individus (RNA-Seq).

MATERIELS ET METHODE :

Aperçu sur les données d'analyse :

Les échantillons de cancer constituaient les marqueurs à prédire tandis que les gènes constituaient les informations caractéristiques à collecter pour la prédiction. On implémente différents scripts qui utilisent des différentes bibliothèques dédier pour le machine Learning, tel que Pandas, Numpy, sklearn et Keras et enfin Matplotlib qui permet de créer et de visualiser les graphiques. Il est recommandé d'installer les bibliothèques dans un environnement Conda pour éviter le problème de dépendance des packages et pour assurer le bon fonctionnement.

Pré-traitement des données :

On réalise une sélection des gènes dont le niveau d'expression présente une caractéristique d'un type de cancer, et une variable prédictive pertinent, et vice versa, afin d'éliminer les gènes qui ne présentent pas de différences d'expression et ne peuvent pas donc nous fournir des informations pertinentes.
On a initialisé l'importation des données d'origine afin de créer un fichier DATA qui contient le numéro des échantillons de 0 à 800 en ligne et les 20 530 gènes en colonne (variables prédictives X). Un autre fichier contient les tumeurs associées à chaque patient, ce qui correspond à la variable Y, celle que l'on souhaite prédite. Les variables prédictives étant trop nombreuses par rapport au nombre d'échantillons, il est nécessaire de trier nos données afin d'éviter le sur ajustement.

Réduction du nombre de caractéristiques (features) :

Afin de rendre les modèles efficace (précision et rappel), la machine doit être parfaitement entraînée pour éviter le phénomène de sous-ajustement (underfitting). Elle ne doit pas apprendre les données qu'on lui a

fournies par cœur, sinon il y a un risque de sur ajustement (overfitting). Dans les deux cas, le pouvoir prédictif du modèle est très faible.

Le choix des algorithmes et l'évaluation :

4 algorithmes de classification ont été appliquer, la régression logistique, l'arbre de décision, le réseau neurone avec Keras et le classificateur extra-arbres.

Concernant la régression logistique, les variables prédictives étant continuées et la variable à prédire catégorique. En revanche, la variable à prédire est multiclasses et non à deux classes, une « simple » régression logistique peut être ne pas adapter à cette analyse.

Un modèle neuronal a ensuite été créé pour utiliser Dropout afin de réduire les risques de sur ajustement. De nombreux nombres de couches, neurones par couches et taux Dropout ont été testés avec peu de précision. Le modèle a été testé pour tous les types de cancer, sur l'ensemble des données. Et 100 epoches ont été nécessaires pour l'achèvement de l'apprentissage (déterminé en atteignant des états sur le graphique de la courbe d'apprentissage). Le réseau de neurones choisi contient 3 couches (une couche d'entrée, une couche de sortie et une couche intermédiaire). Pour la couche intermédiaire, la fonction Relu est utilisée. Elle permet de résoudre les problèmes de saturation des courbes sigmoïde et tangente hyperbolique. Pour la couche de sortie la fonction softmax appliquer, cette fonction est spécialisée dans la classification en multiclasses et est donc adaptée à notre analyse, les valeurs d'entrées sont converties en vecteurs de probabilités dont la somme est égale à 1. Comme fonction de coût à optimiser, en utiliser la fonction « binary_crossentropy ». Elle est adaptée aux cas où la variable à prédire ne peut appartenir qu'à une seule classe et où elle est encodée one-shot.

Par rapport au classificateur extra-arbres, on regroupe les résultats de plusieurs arbres de décision décorrélés collectés dans une « forêt » pour produire son résultat de classification.

Pour les 4 algorithmes citer ci-dessus, en mesuré la capacité prédictive de chaque modèle à l'aide de la fonction de précision afin d'évaluer la performance de la prédiction sont présenté en tableau (figure 4).

RÉSULTATS :

Cet ensemble de données fait partie de données « RNA-Seq (HiSeq) PANCAN ». Il s'agit d'une extraction aléatoire de 20531 expressions génétiques à partir de 801 échantillons de différents patients atteints de tumeurs. On a respectivement 300, 146, 78, 141 et 136 échantillons pour les types de cancer BRCA, KIRC, COAD, LUAD et PRAD (Figure 1). L'ensemble de données des échantillons dans des colonnes. Les niveaux d'expression de chaque gène chez 801 patients sur les lignes avec les différents types de tumeurs associés à chaque patient. En revanche, ce sont des variables numériques continues.

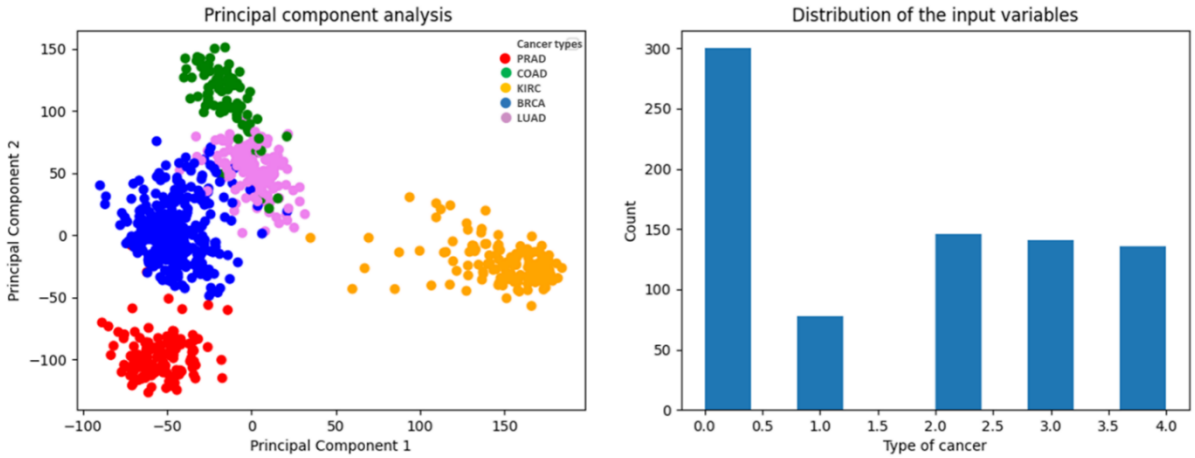

Figure 1 : Analyse prétraitement des données ; Distribution des données d'entrée (A droit). Analyse en Composantes Principales (ACP) des différents de cancers (À gauche).

Selon les résultats de l'ACP, les données sont bien clustérisées, des groupes spécifiques avec des points pour chaque type de cancer. 3 types dans cancers ont des points en commun sont généralement des gènes partager. L'ACP, trouver n'est pas homogène et bien évidement selon les gènes (features) impliqué pour chaque cancer, on a des clusters séparer (figure 1).
Pour le modèle d'arbre de décision, la génération d'une matrice de confusion pour mesurer la performance de classification. On peut observer que le score de précision et supérieur à 0,90 ainsi que chaque type de cancer prédit correspond au type de cancer réel, avec aucune erreur de prédiction (Figure 2).
Le modèle de régression logistique a donné une précision proche de 1 (0,99621) sur le jeu de test et pareil pour le jeu de d'entrainement (figure 3).

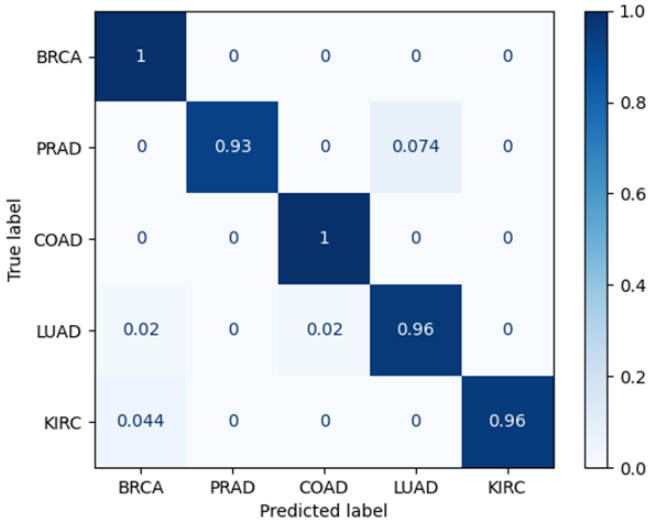

Figure 2 : Evaluation du modèle ; Dotplot à gauche de matrice de confusion du modèle l'arbre de décision.

Une fois notre modèle défini et entraîné, on a évalué son efficacité au travers d'une courbe de précision générée à partir des données prédites, X_train, et de la réalité, X_test. Avant de réaliser l'analyse avec l'architecture définitive, on a réalisé l'analyse, on réduit le nombre des neurones afin d'éviter l'overfitting à l'avance.
Les résultats de la courbe d'apprentissage de réseau neurone qui compare la précision des scores de test et les scores d'entraînement, à montrer que le modèle donnant la meilleure précision en fonction du fil d'expérience (epoch) ce qui suggèrent que la prédiction selon les paramètres choisis sont bien optimisé, on n'a que des vrais positifs et donc une précision proche de 1 (Figure 3) ce qui indique que l'ensemble de données d'entraînement et de test a été parfaitement appris et qu'aucune erreur n'a été commise.
Pour le graphe a droit de la figure 4, qui présente les deux courbes de la fonction de perte (loss function) qui est utilisée pour évaluer une solution candidate (c'est-à-dire un ensemble de poids) afin de minimiser l'erreur. On remarque qu'on n'a pas une grande différence pour les deux courbes sont presque superposées ça renforce l'idée de la précision de modèle choisi.

Types d'algorhitme	Régression logistique	Arbre de décision	ANN (Réseau neurone)	Classificateur d'arbres extrêmement aléatoire (Extra-tree classifier)
Precision-Recall score (données de teste)	0,99621	0,9773	0,999	0,9919

Figure 3 : Tableau des résultats de score de précision pour chaque algorithme machine Learning

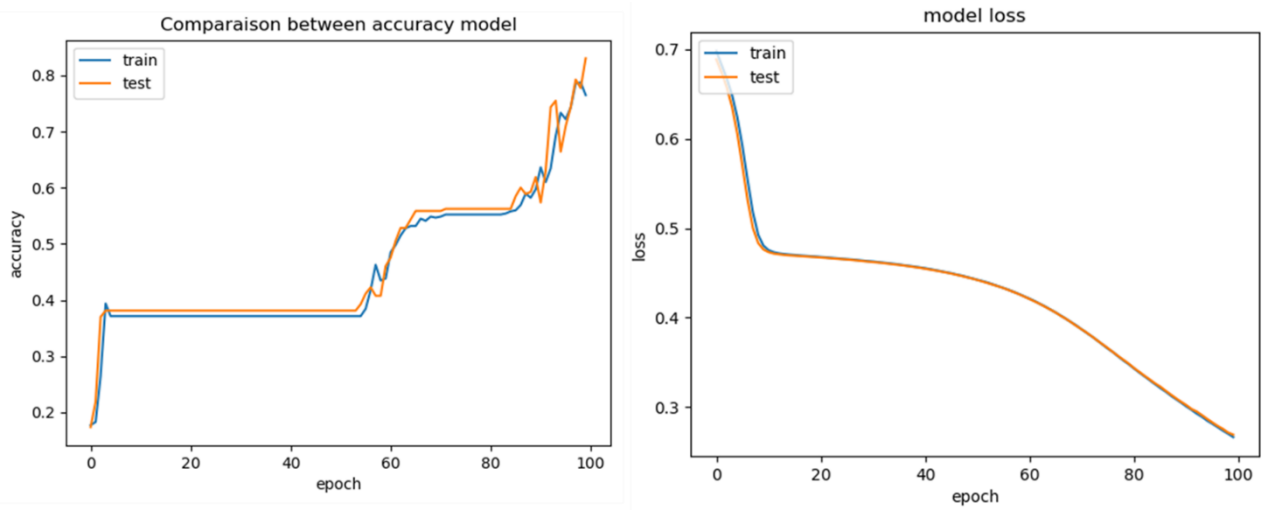

Figure 4 : La courbe représentant l'entraînement (training) est en bleu, celle représentant la prédiction (testing) est en orange. La représentation à gauche représente la précision entre les données d'entraînement et test. Le graphe à droite représente la fonction de perte entre les deux données.

DISCUSSION

Selon l'évolution de l'écart entre les prédictions réalisées par le réseau de neurones et les valeurs réelles des données utilisées pendant la classification. On remarque que les résultats de cette fonction sont minimisés (la courbe diminuer des deux données), donc le réseau de neurones est performant (figure 4).
Concernant le modèle de régression logistique, une précision de 1 sur le jeu de test a été trouvée. Les courbes d'apprentissage ayant des scores élevés, le modèle ne semble pas souffrir d'underfitting. À partir des données, ajouter des données n'est plus utile, car la précision devient fixe. La précision du modèle ainsi que l'allure de la courbe de précision restent étonnantes parce qu'il y a une différence entre les deux données d'entraînement et de test, le score d'entraînement devrait être élevé et le score de test faible.
Selon le modèle de classificateur extra-arbres, on a déterminé les meilleures features (gènes) dans la méthode du classificateur d'arbre supplémentaire, avec la meilleure impureté, c'est-à-dire le moins d'hétérogénéité possible après avoir divisé un groupe d'échantillons en deux groupes à chaque nœud de l'arbre, 10 premiers gènes les plus importants d'après la méthode d'arbres de décision sont : [14115, 19297, 7793, 2775, 2640, 13819, 15896, 15634, 9185, 11904] sont des gènes qui peuvent avoir des rôles principaux dans le cancer.
La construction de notre modèle semble avoir moins prédit les classes de cancer à partir des caractéristiques des données et ça peut être revient à la réduction de dimensions de la variable de nombre de gènes, une réduction bien précise des données d'entrée peut améliorer la prédiction.
De la même façon, une analyse en composantes principales (ACP) peut aussi réduire le nombre de prédicteurs en regroupant les variables associées (covariance) en plusieurs variables principaux composants. Le but est de ne sélectionner qu'une petite quantité de principaux composants afin d'attribuer des individus en fonction de ces composants similaires à une distribution individuelle basée sur tous les prédicteurs.

CONCLUSION :

Dans l'ensemble, l'apprentissage automatique peut être un outil efficace pour prédire des variables à partir d'une grande quantité de données de même nature ou de plusieurs omiques. Ainsi, ce procédé peut ensuite être développé et utilisé dans le cadre de projets de médecine personnalisée pour guider le traitement des patients cancéreux ou avec de maladies génétiques et organiser le dépistage.