Noms : Alain Salanié, Mathias Deroche

Groupe : Les docteurs

Sujet Big Data & Santé (société générale)

Projet : Prévoir le cancer avec les mathématiques

Vous connaissez quelqu'un qui aura le cancer. Peut-être qu'un de vos proches l'a déjà, peut-être que c'est vous qui allez le développer. Il est estimé qu'environ 20 % des gens auront un cancer au cours de leur vie. C'est la première cause de mortalité en France, et aux Etats-Unis, en 2020 environ deux fois plus d'individus sont morts du cancer que du Covid-19.

Devant ces chiffres vertigineux, il faut voir ce que nous pouvons faire. En réalité, un cancer peut se soigner. Généralement les cancers les plus communs (sein pour les femmes, prostate pour les hommes) ont également les taux de survie les plus élevés, dépassant les 90 % après 5 ans. Dans tous les cas, se faire dépister est très important. On estime que 46 % des décès auraient pu être évités si détectés plus tôt. Nous savons ce qui favorise l'apparition de cancer : tabac, alcool, stress... Mais sans dépistage, impossible d'être sûr, car il y a un facteur de génétique, et de chance. Deux vrais jumeaux vivant de la même manière n'auront pas la même expérience avec le cancer. Mais est-il possible de faire une estimation ? De deviner, à partir de données sur la vie des individus, qui est le plus à risque d'avoir le cancer ? C'est ce que nous avons cherché à développer.

Notre projet est donc une intelligence artificielle qui, entraînée sur des données de patient d'hôpitaux, est capable de prédire la propension d'un futur patient à avoir un cancer à l'instant T. Les données récupérées à l'heure actuelle sont : l'âge, le sexe, la fréquence à laquelle les patients fument notée en cigarettes par jour, l'IMC et la consommation d'alcool. Pour les données qui entraînent notre modèle actuel, il est nécessaire d'étiqueter ces données en indiquant si le patient est actuellement diagnostiqué d'un cancer ou non.

A terme nous aurions voulu créer un modèle capable de déterminer à partir de quel âge il devient probable d'avoir un cancer pour les individus, où bien un modèle qui pouvait déterminer si l'on risquait d'avoir un cancer au cours de toute sa vie. Mais pour le premier modèle, cela demanderait de voir l'âge de diagnostic du cancer, plutôt que la simple présence de celui-ci, et pour le second, cela demanderait des données exclusivement sur des gens morts.

Soyons honnête, notre modèle a des limites. Il ne permet pas de bien prendre en compte les changements de mode de vie. Ça arrive que quelqu'un maigrisse, s'arrête de fumer ou se mette à boire à partir de 40 ans, et ça n'aura pas le même impact que quelqu'un qui a ces habitudes depuis toujours. Mais selon nos données, si. Il a cependant une MSE de 0.33, ce qui est plutôt bon pour un si faible nombre de données.

Notre modèle est créé en Python, vous pouvez trouver son code source dans le fichier main.py situé dans le repository github. Le modèle a également été exporté via joblib dans le fichier CancerModelDocteurs.joblib. Vous pourrez trouver notre business canvas dans le fichier Canvas GGh2022.png. Nos données sont dans le fichier Cancerstats.csv.

Nous espérons que vous apprécierez notre projet, pensez à éviter l'alcool, la cigarette et à vous faire dépister fréquemment des cancers les plus communs pour votre sexe.