# Discutez également de ce qui manque dans vos données.

On a constaté après des recherches poussées que le risque d’AVC dépend non seulement de nos variables à expliquer, mais également d’autres variables telles que **la fibrillation auriculaire, l’excès de cholestérol etc.**

# Quelles informations additionnelles voyez-vous comme nécessaire ou utile à avoir pour réaliser votre projet mais manquent au jeu de données.

Les informations additionnelles qui peuvent nous manquer dans notre jeu de données sont de savoir :

* Si la personne a une fibrillation musculaire (une forme d’arythmie cardiaque)
* Si la personne a déjà eu un accident ischémique transitoire (mini-AVC) ou un AVC ;
* Si elle qui souffre de migraines ;
* Si la personne est atteinte d’apnée du sommeil. L’apnée peut entraîner une hausse de la pression sanguine et contribuer à la formation de caillots sanguins ;
* Si la personne a un nombre élevé de globules rouges dans le sang (polyglobulie) ;
* Si la personne a un proche parent a été atteint d’un AVC ;
* Si la personne est atteinte d’un trouble cardiaque (anomalie d’une valve cardiaque, insuffisance cardiaque ou arythmie cardiaque) et celles qui ont récemment eu un infarctus du myocarde.

# Où et comment pourriez-vous obtenir ces informations supplémentaires afin d’enrichir votre jeu de données ?

Pour obtenir ces données on pourrait organiser des enquêtes afin d’en obtenir le maximum pour améliorer notre modèle. Cependant cette méthode serait coûteuse et la collecte serait également difficile parce que souvent les données liées à la santé sont sensibles.

On peut également essayer d’approcher des spécialistes de ce domaine et leur proposer de travailler ensemble dans ce projet et à la fin ils pourront bénéficier de notre modèle final.

…

# Discutez également comment les différents jeux de données interagissent ensemble dans votre système.

Nos différents jeux de données sont structurés à travers des variables dont certains sont corrélées en elles et d’autres non. Cependant il n’y a pas de corrélations entre les variables quantitatives.

Certaines variables comme l’âge sont très liées avec notre variable ***target*** et d’autres n’apportent aucune information nécessaire pour la prédiction de l’AVC.

# Quelles sont les valeurs en commun pour lier les données ensemble ? Comment allez-vous utiliser les données ?

Les valeurs en commun pour lier les variables peuvent être la corrélation qui existe entre certaines d’entre eux. Nous avons constaté par exemple après analyse que la variable **ever\_married** avait une forte corrélation avec celles **age** et **bmi**. Il existe également d’autres variables corrélées entre elles.

Pour leurs utilisations, on a en supprimé certains dès qu’elles avaient une corrélation considérable et en garder celles qui en rapportaient le plus d’informer pour avoir un model optimal.

# Comment allez-vous structurer les données, les stocker, les charger en mémoire ?

Les données ont été mises dans un fichier csv et elles étaient stockée sur *kaggle* et en local. Pour la mémoire, tout s’est passé également sur *kaggle* (la totalité de notre code était sur le cloud).

# Discutez également de la procédure de tests que vous envisagez.

[…]