# STT 3795 - Proposition

**TITRE** : Analyse et prédiction de la maladie cardiovasculaire

**ÉQUIPE** :


- Ronnie Liu
- Si Da Li
- Van Nam Vu


## MOTIVATION ET OBJECTIF DU PROJET :

Les maladies cardiovasculaires sont reliées avec le cœur ainsi que les vaisseaux sanguins. Cet ensemble de maladies est composé de plusieurs terminologies différentes dépendamment de l'organe alimenté par les vaisseaux sanguins. Par exemple, quand on se réfère aux vaisseaux sanguins qui alimentent le cerveau, il s'agit de maladies cérébro-vasculaires. Cependant, lorsque ces derniers alimentent le muscle cardiaque, il s'agit de cardiopathies coronariennes.

Les maladies cardiovasculaires sont « la première cause de mortalité dans le monde […], soit 31% de la mortalité mondiale totale ».[1] Il est donc essentiel de savoir comment prévenir ce type de maladie chez les patients afin de diminuer le taux de mortalité. Il existe déjà des méthodes diagnostiques qui tiennent en compte de facteurs comportementaux dans le but d'évaluer les maladies cardiovasculaires. En effet, l'influence du tabagisme, de la forte consommation d'alcool, de l'obésité, ainsi que de la malnutrition sont seulement quelques facteurs qui peuvent nous donner des indications sur la présence ou non de ces maladies. Cependant, ces méthodes ne sont pas suffisantes étant donné que la liste des facteurs comportementaux est longue et que le temps est limité quant aux observations des patients.

Pour ces raisons, au lieu de se fier uniquement aux facteurs comportementaux, nous avons décidé de créer un modèle optimal permettant de prédire la présence des maladies cardiovasculaires chez différents individus à l'aide de l'apprentissage automatique.

---

[1] Maladies cardiovasculaires. (2017). WHO. https://www.who.int/fr/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)#:~:text=Les%20maladies%20cardiovasculaires%20constituent%20un,sanguins%20qui%20alimentent%20le%20cerveau) [page consultée le 24 février 2021]

**DESCRIPTION DU PROJET :**

On juge qu'un modèle est optimal lorsqu'on obtient une erreur la moins élevée par rapport aux données de test (ou de validation). L'ensemble de données utilisé contient des étiquettes 0 ou 1 (présence ou non des maladies cardiovasculaires chez les patients), donc on fait une classification binaire dans ce projet.

Tout d'abord, nous suggérons plusieurs approches de pré-traitement avant de passer à la phase d'entraînement. Par exemple, on peut réduire la dimension des attributs à l'aide de PCA ou d'Isomap (apprentissage non supervisé). Ensuite, pour chaque ensemble de données qui est prétraité différemment, on évalue la qualité des classifieurs binaires « SVM » et « Forêt Aléatoire » (apprentissage supervisé). Finalement, on prend le meilleur classifieur avec le meilleur pré-traitement comme notre modèle optimal.

En d'autres mots, si on compare la qualité des deux types de classifieurs selon différents ensembles de données (données brutes, traitées sous PCA, traitées sous Isomap), on aura six modèles différents à comparer.

**RÉPARTITION DES TÂCHES :**

- Les tâches seront distribuées de façon équitable entre les membres de l'équipe.
  - Une personne se chargera de la partie d'apprentissage non supervisé: PCA, MDS, ISOMAP, etc.
  - Une personne se chargera de la partie SVM
  - Une personne se chargera de la partie Forêt aléatoire
  - La rédaction du rapport sera faite par les 3 membres de l'équipe.

**JEU DE DONNÉES :**

- Source de données : https://www.kaggle.com/yasserh/heart-disease-dataset
- C'est un jeu de données avec 13 attributs sur 303 patients qui sont les informations et les indices des patients. La dernière colonne est constituée des valeurs 0 (non-présence de la maladie cardiovasculaire) ou 1 (présence de la maladie).

| | |
|---|---|
| 1. **Age**<br>2. **Sex**<br>3. Chest pain type (**cp**)<br>4. Resting blood pressure (**trestbps**)<br>5. Serum cholesterol (**chol**)<br>6. Fasting blood sugar (**fbs**)<br>7. Resting Electrocardiographic (**restecg**)<br>8. Maximum heart rate achieved (**thalach**)<br>9. Exercise induced angina (**exang**) | 10. ST depression induced by exercise relative to rest (**oldpeak**)<br>11. Slope of the peak exercise ST segment (**slope**)<br>12. Number of major vessels (0-3) colored by fluoroscopy (**ca**)<br>13. Displays the thalassemia (**thal**)<br>14. Diagnosis of heart disease (**target**) |

- Informations supplémentaires par rapport à l'ensemble de données : https://archive.ics.uci.edu/ml/datasets/heart+disease