Projet d'étude sur l'espérance de vie

INSA Toulouse, ModIA
UFs "Analyse de données" & "Eléments de modélisation statistique"
Olivier Roustant & Cathy Maugis-Rabusseau

22 septembre 2021

Organisation du projet et documents à rendre

- Le projet sera réalisé par groupe de 4 (ou 3) étudiant-e-s.La constitution des groupes sera faite lors de la première séance. Tous les groupes doivent être mixtes / l'Ecole d'origine (INSA, ENSEEIHT) et si possible / genre (Femme, Homme).
- 8 séances de 2h30 sont dédiées dans votre emploi du temps au travail du projet. Un-e intervenant-e d'une des deux UF sera présent-e lors de ces séances pour répondre à vos questions.
- Livrables : vous devrez rendre (en déposant sous Moodle) au plus tard le **vendredi 28 janvier 2022 minuit** les 3 documents suivants :
  1. un fichier Rmarkdown (*nom1-nom2-nom3-Rapport.Rmd*) contenant les codes R et générant le rapport au format pdf
  2. un rapport au format .pdf (*nom1-nom2-nom3-Rapport.pdf*) généré par la compilation du fichier .Rmd précédent.
     Attention : le rapport est limité à 20 pages, figures incluses.
  3. un notebook jupyter contenant les codes en Python commentés
- Un dossier "ModeleRapport", disponible sur Moodle, vous donne un exemple avec des consignes pour la rédaction de votre rapport. Il est important d'en prendre connaissance dès la première séance !

Evaluation du projet

Pour chaque UF, la note de projet compte pour un tiers de la note finale de l'UF. Elle sera issue de l'évaluation des critères suivants :

Critère	**UF EMS**	**UF AD**
Choix des modélisations ML et MLG adaptées à la question traitée	x	
Utilisation pertinente des méthodes d'exploration de données et de modélisation vues en cours		x
Ecriture mathématique des modèles considérés	x	
Utilisation de méthodes de sélection de variables	x	
Justification des choix de modélisation, des valeurs des paramètres		x
Aller-retour exploration $\leftrightarrow$ modélisation		x
Analyse ($\neq$ lecture !) des résultats obtenus	x	x
Choix et rendu des graphiques illustratifs	x	x
Rédaction d'un document en Rmarkdown	x	
Programmation en R	x	
Couverture du code Python / code R		x
Rédaction générale du document	x	x
Bonus pour des choix originaux adaptés	x	x

Jeu de données étudié

Les données sont issues du référentiel de données de l'Observatoire mondial de la santé (GHO) de l'Organisation mondiale de la santé (OMS). On dispose pour 133 pays de l'espérance de vie, de facteurs de vaccination, de facteurs de mortalité, de facteurs économiques, de facteurs sociaux, collectés de 2000 à 2014. Le jeu de données **LifeExpectationData-Etudiants.csv** est composé des 21 colonnes suivantes :

Nom de la variable	**Signification**
Country	Pays
Year	Année considérée
Status	Statut du pays (Developed or Developing)
Life.expectancy	Espérance de vie en année
Adult mortality	Taux de mortalité des adultes des deux sexes (probabilité de mourir entre 15 et 60 ans pour 1000 habitants)
Infant deaths	Nombre de décès infantiles pour 1000 habitants
under.five.deaths	Nombre de décès d'enfants de moins de cinq ans pour 1000 habitants
HIV.AIDS	Nombre de décès pour 1 000 naissances VIH/SIDA (0-4 ans)
Alcohol	Consommation d'alcool enregistrée par habitant (15+) (en litres d'alcool pur)
BMI	Indice de masse corporelle moyen de l'ensemble de la population
thinness..1.19.years	Prévalence de la maigreur chez les enfants et adolescents de 10 à 19 ans (%)
thinness..5.9.years	Prévalence de la maigreur chez les enfants et adolescents de 5 à 9 ans (%)
Measles	Nombre de cas de rougeole signalés pour 1000 habitants
percentage.expenditure	Dépenses de santé en pourcentage du produit intérieur brut par habitant (%)
Total.expenditure	Dépenses publiques générales de santé en pourcentage des dépenses publiques totales (%)
Hepatitis.B	Couverture vaccinale contre l'hépatite B chez les enfants de 1 an (%)
Polio	Couverture vaccinale contre la polio chez les enfants de 1 an (%)
Diphtheria	Couverture vaccinale contre l'anatoxine diphtérique, le tétanos et la coqueluche chez les enfants de 1 an (%)
GDP (PIB)	Produit Intérieur Brut par habitant (en USD)
Income.composition.of.resources	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
Schooling	Nombre d'années de scolarité (années)
Population	Population du pays

Questions à aborder

Dans votre rapport final, vous devez avoir abordé par une/des méthodes adaptées les questions suivantes :

- Analyse du jeu de données. Préparation du jeu de données (variables redondantes ? transformations ? création de features ? outliers ?...) et visualisation (dans un espace de faible dimension).
- Etude de l'espérance de vie en 2014 :
  - Comment les taux de mortalité infantile et adulte affectent-ils l'espérance de vie en 2014 ? Même question si on considère en plus le statut du pays.
  - Quel est l'impact de la couverture vaccinale sur l'espérance de vie en 2014 ?
  - Quelles sont parmi toutes les variables prédictives celles expliquant l'espérance de vie en 2014 ?
- On considère maintenant toutes les espérances de vie (la variable année peut être considérée comme une variable explicative).
  - Quelles sont parmi toutes les variables prédictives celles affectant réellement l'espérance de vie ?
  - Quelles sont les variables prédictives qui permettent de discriminer entre une espérance de vie inférieure ou supérieure à 65 ans ?
  - Même question pour discriminer entre une espérance de vie inférieure à 65 ans, entre 65 et 80 ans, ou plus de 80 ans ?
- Pour chaque cas (régression, classification), comparer en terme de capacité prédictive les modèles linéaires avec les modèles non-linéaires (arbres, forêts aléatoires).
- Voit-on des clusters dans les données ? Ceux-ci aident-ils à comprendre les différences d'espérance de vie ?