Examen

Rindra LUTZ

27/01/2021

Travaux des autres groupes

Rattle

Lien du travail à l'adresse suivante "https://github.com/Nicolas-all/PSB1"
Travail effectué par **Nicolas ALLIX** et **William ROBACHE**

Ce travail a été abordé de manière à ce qu'un lecteur peu connaisseur du sujet puisse s'y retrouvé, tout en étant un travail assez complet pour prendre en main le package Rattle et nombre de ses fonctionnalités.
Rattle est utilisé en R pour réaliser du data mining. Rattle présente des résumés statistiques et visuels des données, et transforme les données qui peuvent être facilement modélisées. Il permet en outre de construire à la fois des modèles non supervisés et supervisés, ce qui le rend polyvalent.
Rattle présente un interêt vis à vis des interactions de l'utilisateur sur le GUI, qui sont enregistrées en script R. Ce script peut ensuite être utilisé et exécuté facilement, indépendamment de l'interface Rattle en elle-même.

Une des commandes les plus intéressantes à utiliser via Rattle est la commande :
evaluateRisk(predicted, actual, risks)

Cette commande va permettre de résumer les performance d'un modèle étudié. En prenant les valeurs prédites, les valeurs réelles et les mesures du risque associées à chaque cas, cette fonction génère une synthèse complète qui regroupe les valeurs prédites distinctes, en calculant le pourcentage cumulé de cas, de rappel, de risque, de précision et de mesure. Cela fait donc de cette commande un des principales commandes à utiliser sur Rattle, puisque lors des études en Data Mining, il est absolument indispensable d'évaluer ses performances vis à vis d'un modèle choisi.
Il existe bien entendu de nombreuses fonctionnalités de code très intéressantes, telle que la randomForest2Rules qui va permettre de parcourir tous les arbres des random forest et générer un ensemble de règles, ou bien le calcul de l'AUC via calculateAUC(x, y), qui va pouvoir nous aider à déterminer l'aire sous une courbe d'un graphique de risque.

Somme toute un travail complet, concis, et qui permet d'apprendre à utiliser simplement beaucoup de fonctionnalités de Rattle. Sont présentes de nombreuses définitions et exemples de codes, ce qui en fait une très bonne base d'études, même pour les utilisateurs les plus novices. C'est donc un bbon travail ouvert à la portée de tous, ce qui en fait un travail à ne pas manquer.

Flexdashboard

Lien du travail à l'adresse suivante "https://github.com/Nicolas-all/PSB1"
Travail effectué par **Nicolas ALLIX** et **Marko ARSIC**

Les tableaux de bord ou "Dashboard" constituent des outils graphiques qui sont utilisés pour répresenter de façon simple les données. C'est un moyen de visualisation permettant de regrouper un ensemble de données très variées, et résidant au coeur des métiers de la data.
On retrouve dans ce travail sur le package Flexdashboard une structure claire, avec une explication des basiques du package, ses composants, et uine introduction à son utilisation avec Shiny. Flexdashboard permet de personnaliser les visualisations de données avec énormément de liberté, à commencer par la disposition des graphiques.

Bien que n'ayant pas de code réellement présent dans le document de travail, il est très simple de trouver ce genre de code. Le code suivant donne un aperçu de l'utilité du package Flexdashboard :

Chart 1

Chart 2

Chart 3

Il est ici tout simplement question de superposer trois parties, mais il aurait été également possible de n'en superposer que deux et d'en mettre une troisième en colonne à droite.
Pour bien comprendre et utiliser ce package de positionnement, rien ne vaut l'exemple, et il est donc préférable de tester différentes combinaisons jusqu'à s'approprier les connaissances du sujet.

Ce travail très pratique et théorique permet donc d'acquérir les outils et la compréhension nécessaire à l'utilisation du package, sans rentrer dans des exemples concrets répétitifs.
C'est donc au lecteur de tester ses connaissances sur l'outil, ce qui n'est pas plus mal popur se faire la main.

Tenserflow

Lien du travail à l'adresse suivante "https://github.com/Nicolas-all/PSB1"
Travail effectué par **Nicolas ALLIX** et **Marko ARSIC**

TensorFlow est une plate-forme d'apprentissage automatisée et open source. Elle fournit un écosystème complet et flexible d'outils, de bibliothèques et de ressources communautaires, permettant aux chercheurs de progresser dans le domaine de l'apprentissage automatisé.
TensorFlow utilise un écosystème complet de machine learning pour résoudre des problèmes complexes, ce qui en fait un incontournable dans le domaine des apprentissages automatisés. L'une des particularités de TensorFlow est également l'utilisation de graphes de flux de données.
L'interface TensorFlow se compose de nœuds, qui représentent des calculs mathématiques, et de tenseurs. Un tenseur peut être un entier, un vecteur, une image… Chaque nœud du

graphe prend donc en entrée différents tenseurs, effectue son calcul, puis retourne de nouveaux tenseurs.
Les liens placés en bas de page de ce travail nous révèle une plus grande complexité qu'à la première impression, le document de travail n'étant pas spécialement long. On a cependant à notre disposition tout un tas d'explications et d'exmples dans ces liens, qui permettent au lecteur de comprendre plus en profondeur le sujet. Ce sujet, assez théorique, nous en apprend plus sur le système de tenseurs:
Le code associé à TensorFlow se divise en deux étapes principales, la construction et l'exécution. Pendant la phase de construction, les variables et les opérations du graphe sont définies et assemblées. La création du graphe est ensuite automatiquement gérée par TensorFlow afin de permettre l'optimisation et la parallélisation du code et de l'exécution.

La phase d'exécution utilise une session afin d'exécuter les opérations du graphe. Un graphe ne va exécuter les opérations qu'après création d'une session. Une session permet de placer les opérations du graphe dans des les composants (CPU/GPU/TPU) et met à disposition des méthodes pour les exécuter. Le lancement des opérations du graphe se fait via la méthode run() de la session comme nous le verrons un peu plus loin. Ce système d'exécution de graphe est une des propriétés fondamentales de TensorFlow et permet d'exécuter toutes les opérations du graphe en une seule fois.

L'absence d'exemple concret sur ce travail est donc largement compensée par la présence de liens ayant une documentation plus poussée, ce qui en fait une introduction naturelle au sujet. Il sera donc grandement efficace de se rapprocher du sujet en passant par ce travail, alliant introduction, théorie et pratique via des liens externes.

ggplot2

Lien du travail à l'adresse suivante "https://github.com/Nicolas-all/PSB1"
Travail effectué par **Nicolas ALLIX** et **Marko ARSIC**

Ggplot2 est un package de visualisation utilisé en R et en python. On l'utilise généralement pour réaliser des graphiques à partir d'une base de données, mais ce package peut largement aller plus loin dans sa complexité.
Il est notamment possible de créer toutes sortes de cartographies pour avoir une visualisation de type google map, afin de faire ressortir les informations et idées d'une base de données de manière concrète par région du monde.
Avec ggplot2, il est donc tout à fait possible de réaliser une carte météorologique, une carte sur la propagation de différentes épidémies ou encore d'étudier des critères sociaux démographiques. L'utilité de ggplot2 est de faire ressortir l'information que l'on veut faire passer de manière visuelle.

```
# Calculer le centroïde comme étant la longitude et la lattitude moyennes
# Utilisé comme coordonnée pour les noms de pays
# region.lab.data <- some.eu.maps %>%
# group_by(region) %>%
# summarise(long = mean(long), lat = mean(lat))
```

Sur cet exemple de code présent dans le travail effectué, on peut se rendre mieux compte de l'importance et de l'utilité de ggplot2. En effet, après avoir récupéré certaines données géographiques, il est maintenant possible de nommer les payes ainsi que de récupérer les données longitudinales et latitudinales. Ces données sont importantes tant en terme de visualisation qu'en terme de précision, car elles apportent un vrai plus quant à la localisation de certaines données de la base étudiée. Il est donc particulièrement intéressant de développer le code ci-dessus afin de répondre avec une précision accrue aux demandes géographiques possibles.

Le travail effectué sur ggplot2 montre à travers un court exemple applicatif l'utilité de ce package. Dans le domaine de la data, la représentation visuelle est un des éléments les plus importants, car les résultats attendus par certains demandeurs doivent être le plus clair possible. Il faut donc qu'ils soient accessibles à tous, et pas seulement aux utilisateurs des logiciels de traitement de données. Un incontournable donc !

Good Prediction

Lien du travail à l'adresse suivante
"https://github.com/Cldren/REN_PSBx/tree/main/Maths"
Travail effectué par **Claude REN**, **Claire MAZZUCATO** et **Thuy AUFRERE**

Ce papier tente de démontrer l'importance de mettre en place des approches d'évaluation sur les prédictions. Pour comparer les prédictions, il est nécessaire de disposer de mesure ou d'un moyen d'évaluation.
Cependant, ce n'est pas parce qu'une prédiction est exacte qu'elle est utile pour informer le comportement. En effet, le degré d'erreur d'une prédiction n'informe pas sur l'utilité d'une prédiction pour informer d'autres prédictions. Pour évaluer une prédiction, il est nécessaire de mettre l'accent sur la pertinence des caractéristiques. Pour les auteurs, une bonne prévision est celle dont les caractéristiques sont bien alignées avec le problème de prédiction.

Ainsi ce papier cherche à montrer que les méthodes actuelles communes d'évaluations des connaissances prédictives ne permettent pas de faire la différence entre les prédictions utiles et les prédictions ordinaires : c'est-à-dire entre les prédictions utiles pour la prise de décision et l'apprentissage et celles qui ne le sont pas. Les auteurs proposent d'améliorer les méthodes actuelles d'évaluations en suggérant aussi d'évaluer les prédictions non seulement en fonction de leur propre erreur de prédiction mais aussi de l'erreur des autres prédictions qui en dépendent.

La formule principale du papier est la formule du GVF, qui va aider à déterminer une prédiction. Les fonctions de valeur stockent la récompense cumulative prévue d'un agent à partir d'un état. Chaque fonction de valeur peut être décrite comme une réponse à une question de départ.

Les GVF peuvent être interprétées comme un rendement attendu à une action C. En ce qui concerne les prédictions, à chaque action C, on obtient un vecteur d'observations qui décrit l'environnement et prend une mesure. Les observations sont utilisées pour construire la fonction ϕ.

Cette fonction à laquelle on applique un poids va être l'estimation de la prédiction grâce à laquelle nous pouvons calculer l'erreur. C'est donc une fonction importante qui va servir de socle à tout le travail en aval, c'est pourquoi elle est importante à retenir dans l'idée de réaliser de bonnes prédictions.

Ce passage dans la théorie mathématiques du domaine de la prédiction en statistiques est vraiment intéressant. En effet, il permet de poser les bases de la prédiction, qui est un thème de plus en plus abordé dans notre société : que vont devenir les chiffres, peut-on prévoir une action, etc… La prédiction figure en thème majeure de notre époque, et la bonne compréhension de la théorie qui précède l'application en entreprise de la prédiction est un véritable plus en soit, un apport supplémentaire dans les capacités d'un statisticien à comprendre et interpréter ses résultats.

Dropout Prediction

Lien du travail à l'adresse suivante
"https://github.com/Cldren/REN_PSBx/tree/main/Maths"
Travail effectué par **Claude REN**, **Claire MAZZUCATO** et **Thuy AUFRERE**

Les cours ouverts en ligne (MOOC) sont des plateformes populaires d'apprentissage en ligne.
L'hypothèse est ici que le clickstream d'un élève sous-entend certains modèles comportementaux d'abandon. Cependant la prédiction d'un comportement d'abandon est complexe. Tout d'abord, il est difficile d'extraire une représentation comportementale significative d'un utilisateur à partir de données de faible niveau sur le flux de clics, car l'ensemble de données dépend de nombreux facteurs différents et peut-être inconnus tels que le style d'apprentissage de l'utilisateur, le programme ou encore le contenu de la semaine. En outre, il est difficile d'utiliser les approches existantes pour analyser les données de flux de clics.

La méthode proposée ici est de construire une représentation significative et efficace du modèle de comportement d'un élève à partir de son flux de clics sur la plateforme. L'objectif est d'obtenir ce modèle de manière non supervisée afin que des modèles de classification simples comme le perceptron multicouche (MLP) puissent prédire l'abandon de manière comparable à des modèles de classification plus complexes.

Les résultats obtenus à l'aide de l'algorithme Branch and Bound contiennent deux types d'informations : les informations séquentielles entre les semaines et les informations sur la cooccurrence au niveau des actions. Pour compléter ce modèle, les auteurs utilisent le perceptron multicouche (appelé MLP) pour générer une représentation des caractéristiques qui saisit les deux informations.

Le but de l'apprentissange est de minimiser cette fonction.
L'équation est une double somme sur l'ensemble des semaines t et sur un intervalle $2w$ autour de la semaine t. La fonction au carré $(\hat{y}_t - x_{t+1})^2$ est utilisée de manière à prendre en compte des écarts positifs. Le facteur $\frac{1}{2wT}$ est un facteur de normalisation.
Cette fonction est intéressante, car elle permet de comprendre sur quoi l'apprentissage va se faire. Afin de la minimiser, on devra avoir un minimum de somme le plus bas possible, et ainsi avoir un apprentissage plus fiable en sortie.

Ce travail est àà la fois instructif et court, ce qui en fait un travail rapide à lire pour finalement une compréhension assez bonne. Le rapport qualité-temps passé à le lire est très rentable, et c'est pourquoi cela en fait un bon travail pour comprendre la théorie qui se cache derrière la qualité de l'apprentissage et de la prédiction en général.

Early Prediction

Lien du travail à l'adresse suivante
"https://github.com/Cldren/REN_PSBx/tree/main/Maths"
Travail effectué par **Claude REN**, **Claire MAZZUCATO** et **Thuy AUFRERE**

Le sujet est ici la prédiction du comportement d'apprentissage des étudiants à risque afin d'intervenir à temps en cas d'abandon scolaire.

L'étude porte ic sur une expérience approfondie sur un ensemble de données à grande échelle. Les traces d'apprentissage en ligne proviennent de la façon dont les étudiants utilisent le Blackboard, une plateforme en ligne d'apprentissage. Ainsi, des données de parcours ont été recueillies avec les horodatages de certains des modules les plus populaires, y compris la connexion, la déconnexion, l'accès au matériel de cours, les devoirs, le forum de discussion, etc.

Cette équation est utilisée pour caractériser l'homophilie. Pour ce faire, on commence par parcourir le voisinage de chaque noeud qui est censé représenter les personnes avec qui le noeud est le plus proche. Ce parcours de noeud est caractérisé par la fonction $p(c_i = u|c_{i-1} = v) = \frac{\alpha_{pq} w_{uv}}{Z}$ lorsque deux étudiants sont allés à la bibliotèque au même moment, et 0 sinon. Dans cette formule, Z est une constante de normalisation et α_{pq} est un coefficient qui varie selon la distance la plus courte entre les deux noeuds.

La méthode d'apprentissage utilisée se base sur la méthode du maximum de vraisemblance. Pour découvrir des caractéristiques statistiques significatives, les auteurs effectuent un test ANOVA afin de déterminer quels comportements sont statistiquement significatifs pour distinguer les élèves à risque des élèves normaux. Ainsi, cette formule mathémtique est à la base des travauxx en Anova, travaux poussés avec lesquels de nombreuses recherches avancées sont réalisées.

Ce travail, comme précédemment, est donc une bonne base concernant la compréhension d'un sujet en particulier, auquel s'ensuit l'aspect pratique, via ici les tests Anova. Il est donc un bon appui sur lequel on peut compter pour comprendre encore un peu mieux les méthodes d'apprentissage et la théorie qui les entoure.

StatsbombR

Lien du travail à l'adresse suivante
"https://github.com/Cldren/REN_PSBx/tree/main/Packages_presentation/Statsbomb"
Travail effectué par **Claude REN**

Le package StatsBombR nous fournit des données libre d'accès à propos du football. Le package contient différentes fonctions qui facilitent l'extraction des données, et ces fonctions sont ce qui est étudié dans ce travail.

Premièrement, nous avons droit à une bonne introduction du package et de sa fonctionnalité.

Après le chargement des données, on peut filtrer de nouveau afin de ne regarder qu'une seule correspondance. On peut avoir un regard sur les données et trouver le match_id que l'on veut étudier. Ou si vous étudiez une base de données plus grande, vous pouvez faire quelque chose de similaire en filtrant par team.name et utiliser la fonction distinct() sur match_id pour extraire tous les identifiants de correspondance de l'équipe que vous avez sélectionnée. La fonction mutate() est utilisée pour écrire NA par 0 dans la colonne xG avec la fonction if_else(condition, true, false).

Le travail effectué ici est plus complet que les travaux précédents, dans le sens où une mise en situation par un exemple peu court est mise en avant. Cela apporte un vrai plus dans la compréhension et l'apprentissage du package, car le lecteur peut réaliser les tests de code en même temps que sa lecture et voir par lui même les actions effectuées. C'est donc un travail pédagogique qui laisse place à la pratique, un très bon point.

Time domaine approache

Lien du travail à l'adresse suivante "https://github.com/Siva-chane/PSBX"
Travail effectué par **Siva CHANEMOUGAM**

Le sujet traité ici est le sujet des fonctions continues et discrètes dans un système dynamique a partir d'un échantillon de données. Plus précisément, il est nécessaire ici d'identifier le modèle continu associé à un modèle discret.

En grande partie orienté sur la méthode traditionnel SVF, il est tout de même mentionné et un peu étudié la méthode stochastique.
La méthode SVF repose sur le principe suivant : les conditions initiales de l'équation différentielle sont connues. Comme les conditions initiales sont connues, on sait estimer les dérivés en certains points.

Le travail se concentre sur deux fonction, en entrée la fonction u(t) et en sortie une fonction y(t) qui réponds à une équation différentielle présente dans le document. L'équation

différentielle est complexe à résoudre, car elle comprend plusieurs dérivées. C'est pour cette raison qu'il y a discrétisation du problème, et ainsi, grâce à différentes méthodes il est possible de résoudre le problème.

La complexité des formules mathématiques explicitées sur le travail effectuée est intéressante. En effet, cela rajoute du challenge au lecteur quant à la compréhension profonde su sujet, et de bonnes relectures sont de mises. Cela ne joue pas forcément en la défaveur du travail, car l'on s'imprègne plus encore dans le sujet, et on comprend mieux l'importance des équations différentielles dans un problème comme celui-ci. Il est en outre important de voir différentes méthodes de résolutions, et c'est ce qui nous est proposé dans ce travail.

Rpart

Lien du travail à l'adresse suivante "https://github.com/Siva-chane/PSBX"
Travail effectué par **Siva CHANEMOUGAM**

Pour ce dernier travail à analyser, parlons du package Rpart.
Le package RPART permet de construire des modèles de classification ou de régression d'une structure très générale en utilisant une procédure en deux étapes. Les modèles résultants peuvent être représentés sous forme d'arbres de decision. Le package met en œuvre de nombreuses idées trouvées dans le livre et les programmes CART de Breiman, Friedman, Olshen et Stone.

Les arbres de décision font partie de la catégorie des algorithmes supervisés, ils permettent de prédire une valeur (prédiction) ou une catégorie (classement). C'est une méthode très populaire en Data Science et qui a donné naissance à d'autres algorithmes tels que le Random Forest. Comme son nom l'indique, cet algorithme se base sur la construction d'un arbre ce qui rend la méthode assez simple à expliquer et plus facile à interpréter.

Les trois grands points positifs des arbres de décision sont : sa facilité de compréhension, sa facilité d'interprétation, et le temps d'exécution desprogrammes assez raisonnable.

La construction d'un arbre de décision se fait selon la commande

```
ptitanic.Arbre <- rpart(survived~.,data=
ptitanic.apprt,control=rpart.control(minsplit=5,cp=0))
```

Cet exemple de code est très important, car toute l'étude faite sur la base de données va se réaliser à partir de l'arbre de décision que cela sortira. La formule utilisée indique qu'on souhaite prédire la variable survived en fonction de toutes les autres. Le principe général est que la (ou les) variable(s) à prédire sont à gauche du symbole ~ alors que les variables prédictives sont à droite du symbole. Ici, le point . permet d'indiquer qu'on souhaite utiliser toutes les variables des données comme prédicteurs sauf les variables à prédire. La commande rpart.control permet préciser les éléments à modifier dans la construction de l'arbre, le minsplit permet de decouper les feuilles qui contiennent au moins 5 observations, cp=0 pour dire sans contrainte de decoupage.

On a affaire ici avec un cas d'exmple concret, qui permet de prendre en main les commandes à appliquer lors de la création d'un arbre de décision. Rpart est donc un package performant, adapté, et accessible à tout type de personnes. Appliquer ses fonctionnalités sur R est assez simple, et c'est un incontournable en matière d'analyses basiques.
Un très bon travail a été fait sur ce package, et ce travail a été approfondi tout en mettant en avant l'aspect pratique via R. C'est un des travaux les plus explicites que chacun se doit de regarder et de comprendre pour mieux intégrer les analyses via les arbres de décision.

Mes travaux

Cross-Validation

Lien du travail à l'adresse suivante "https://github.com/rindra-lutz/psb1/tree/main"
Travail effectué par **Rindra LUTZ**

La cross validation est une régression logistique, qui font parties des méthodes prédictives. Elle sert à mesurer la fiabilité d'un modèle de prédiction étudié, et est ainsi une étape primordiale dans la validation d'un projet en cours.
Il existe plusieurs méthodes de cross validation dont :
• LOOCV (leave-one-out cross-validation)
• LKOCV (leave-k-out cross-validation)
• k-fold cross-validation

Ces trois méthodes sont les plus utilisées, et les plus fiables dans la plupart des cas. La principale est la k-fold cross validation, largement étudiée et utilisée, que ce soit au niveau postbac ou en entreprise.
Pour réaliser une bonne cross validation, il est nécessaire de diviser son échantillon de données en plusieurs sous-échantillons de taille similaires, afin de réaliser des phases d'entrainements et de tests sur chacun des sous-ensembles tour à tour.
En clair, les étapes sont les suivantes :
1- Diviser aléatoirement l'ensemble de données en k sous-ensembles (par exemple 5 sous-ensembles)
2- Mettre de côté un sous-ensemble et former le modèle sur tous les autres sous-ensembles
3- Tester le modèle sur le sous-ensemble mis de côté et enregistrer l'erreur de prédiction
4- Répétez ce processus jusqu'à ce que chacun des k sous-ensembles ait servi d'ensemble de test.
5- Calculez la moyenne des k erreurs enregistrées. C'est ce qu'on appelle l'erreur de validation croisée qui sert de mesure de performance pour le modèle.

```
# Définiiton de l'échantillon d'entraînement
# set.seed(123)
# train.control <- trainControl(method = "repeatedcv",
#                               number = 10, repeats = 3)
# Entraîner le modèle
# model <- train(Fertility ~., data = swiss, method = "lm",
#                trControl = train.control)
```

```
# Résultats résumés
# print(model)
```

Le code ci-dessus est particulièrement intéressant. En effet, il s'agit d'une étape importante dans la réalisation d'une cross validation : c'est la phase d'entraînement. Cet entraînement va permettre d'estimer la précision du modèle choisi, et ainsi valider ou non ce modèle. Si la validation n'est pas acceptée, il faudra alors choisir un autre modèle de prédiction.
On comprend donc mieux pourquoi la validation croisée est utilisée : il est bien mieux de se rendre compte de ses erreurs et les rectifier quitte à y passer du temps que de fournir un modèle de prédiction totalement erronné, et ainsi nuire à la stratégie de son entreprise.

FactoMineR

Lien du travail à l'adresse suivante "https://github.com/rindra-lutz/psb1/tree/main"
Travail effectué par **Rindra LUTZ**

FactoMineR est un package R dédié à l'analyse exploratoire multidimensionnelle de données.

Pourquoi FactoMineR ?

- Il permet de mettre en oeuvre des méthodes analyses de données très utilisées telles que l'analyse en composantes principales (ACP), l'analyse des correspondances (AC), l'analyse des correspondances multiples (ACM), ainsi que des analyses plus avancées.

- Il permet l'ajout d'informations supplémentaires telles que des individus et/ou des variables supplémentaires.

- Il fournit un point de vue géométrique et de nombreuses sorties graphiques.

- Il fournit de nombreuses aides à l'interprétation (description automatique des axes, nombreux indicateurs, ...).

- Il peut prendre en compte diverses structures sur les données (structure sur les variables, hiérarchie sur les variables, structure sur les individus).

- Beaucoup de matériels pédagogique (MOOC, livres, etc.) est disponible pour expliquer aussi bien les méthodes que la façon de les mettre en oeuvre avec FactoMineR.

- Il gère les données manquantes avec missMDA.

- Il a une interface Shiny qui permet de construire des graphes de façon interactive avec Factoshiny.

- Il propose une interprétation automatique des résultats obtenus avec FactoMineR grâce à FactoInvestigate.

FactoMineR est très utilisé en ce qui concerne les ACP.
Afin d'avoir une meilleure idée des données, il peut être utile de visualiser les valeurs propres. Le code ci-dessous permet de montrer le pourcentage de variances expliquées par chaque axe principal.

```
#eig.val <- res.pca$eig
#barplot(eig.val[, 2],
#        names.arg = 1:nrow(eig.val),
#        main = "Variances Explained by PCs (%)",
#        xlab = "Principal Components",
#        ylab = "Percentage of variances",
#        col ="steelblue")
# Add connected line segments to the plot
#lines(x = 1:nrow(eig.val), eig.val[, 2],
#      type = "b", pch = 19, col = "red")
```

Ce code est une des clés permettant de réaliser une bonne ACP. En effet, les valeurs propres apportent un véritable indice concernant la qualité d'un modèle étudié, et sont ce sur quoi se reposent nombres de statisticiens en amont de leur travaux pour définir cette qualité.

FactoMineR est donc un package complet, fait pour la prise de décision suite à un projet. Il possède de nombreuses fonctionnalités et permet d'apporter au statisticien et à l'entreprise une précision supplémentaire sur les modèles mis en avant, et ainsi aider à la prise de décision.

Comparaison des travaux

Tous les travaux que j'ai été amené à voir sont d'une excellente facture. Tous les étudiants ont joué le jeu et leur sérieux s'est fait ressentir sur la qualité de leurs travaux. En comparant avec mes travaux, on retrouve beaucoup de similitudes dans la manière d'écrire, avec une introduction, des définitions et une approche théorique, puis, quand cela est possible, une approche plus pratique avec du code R. Sur certains travaux, les exemples n'étaient peut être pas très nombreux, mais cela n'influt pas spécialement sur la qualité. En effet, là où le texte ne paraissait pas très long, il a été possible de s'orienter sur le sujet via des liens externes. Ainsi, toutes les informations nécessaires à la compréhension de chaque sujet pouvaient se retrouver facilement. C'est donc avec plaisir que les lectures des différents travaux se faisaient !
On peut également noter que les autres approches concernant les mêmes travaux que j'ai réalisé étaient diverses. Bien que je n'en ai pas parlé dans ce fichier, il est donc positif de voir différents points de vue et chemins pour expliquer un domaine précis de la statistique ou des mathématiques en général.
Concrètement parlant, les travaux sont donc très bons dans l'ensemble et n'ont pas grand choses envier les uns aux autres. Un peu plus de code pour une compréhension pratique plus optimale aurait pu être un plus dans certains de ces travaux, mais comme dit précédemment, on peut retrouver des exemples facilement en prenant le temps de regarder les liens externes mis à notre disposition.
Finalement, un très bon travail de la part de toute la promotion !