Jeu de données 3

Variables

Droits_de_lHomme:
The scores capture the extent to which citizens' physical integrity is protected from government killings, torture,
political imprisonments, extrajudicial executions, mass killings and disappearances. Higher scores mean fewer
such abuses. A score of 0 means an average score relative to all countries and years. A score of 2 is two standard deviations better than the mean across all countries and years.

Liberte-economique 0 à 10 (10 best)
The index measures the degree of economic freedom present in five major areas: [1] Size of Government; [2] Legal System and Security of Property Rights; [3] Sound Money; [4] Freedom to Trade Internationally; [5] Regulation. Within the five major areas, there are 42 distinct variables, which are averaged to derive the summary rating for each country. Scores are on a scale of 0-10, where 10 represents maximum economic freedom.

Inegalite_des_genres 0 à 1 (0 moins d'inégalités)
The GII is an inequality index. It shows the loss in potential human development due to disparity between female and male achievements in two dimensions, empowerment and economic status. The GII ranges between 0 and 1. Higher GII values indicate higher inequalities and thus higher loss to human development.
(Critiqué par certains auteurs mais utilisé par les nations unies)

Proportion_denfants_harceles
Percentage of children aged 13-15 who reported being bullied at least once in the past couple of months.

Corruption (100 (très propre) à 0 (très corrompu)
Annual ranking of countries by their perceived levels of corruption, as determined by expert assessments and opinion surveys. Scale is from 100 (very clean) to 0 (highly corrupt).

Taux_alphabetisation
Estimates correspond to the share of the population older than 14 years that is able to read and write.

HDI
The Human Development Index (HDI) is a summary measure of key dimensions of human development: a long and healthy life, a good education, and having a decent standard of living.

Annees_detudes
Average total years of schooling for adult population

Esperance_vie

Life expectancy at birth is defined as the average number of years that a newborn could expect to live if he or she were to pass through life subject to the age-specific mortality rates of a given period.

Budget_militaire_from_GPD
Military expenditure as a share of GPD

GPD = PIB (permet de quantifier la valeur totale de la « production de richesse » annuelle effectuée par les agents économiques (ménages, entreprises, administrations publiques) résidant à l'intérieur d'un territoire.)

Problème avec ces données aussi. Elles ont été mises à jour.

Sous_alimentation
Share of people of the population who are undernourished
Undernourishment measures the share of the population that has a caloric intake which is insufficient to meet the minimum energy requirements necessary for a given individual.

Acces_Internet_pers
Share of the population using the Internet

Acces_electricite_pers
The percentage of population with access to electricity

Pop_urban & Pop_rural (nombre)
Number of individuals living in cities or not

Analyse descriptive

Analyse en composantes principales (ACP)

Clustering

Ajout de la variable Gscore

Modèle linéaire

Nous avons tout d'abord cherché à expliquer la variable Gscore en fonction d'une unique variable

Par la suite, nous avons voulu expliquer la variable Gscore en fonction de toutes les autres variables. On avait un R^2 de 0.75 donc nous pouvons en déduire que le niveau de démocratie du pays est plutôt bien expliqué avec les variables du jeu de données 3.

```
Call:
lm(formula = Gscore ~ ., data = data3)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1709 -0.5898  0.0628  0.7134  2.9385

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -2.440e-01  2.649e+00  -0.092 0.926737
Droit_de_lHomme              1.595e-01  1.141e-01   1.397 0.164720
Liberte_economique           7.793e-02  1.759e-01   0.443 0.658431
Inegalite_des_genres         2.132e+00  1.344e+00   1.586 0.115138
Proportion_denfants_harceles -7.484e-03  1.005e-02  -0.745 0.457684
Corruption                   5.179e-02  1.304e-02   3.973 0.000117 ***
Taux_alphabetisation        -1.101e-02  1.152e-02  -0.956 0.340924
HDI                          8.182e+00  3.511e+00   2.330 0.021319 *
Annes_detudes               -5.236e-04  9.710e-02  -0.005 0.995705
Espérance_vie               -7.498e-03  3.768e-02  -0.199 0.842577
Budget_militaire_from_GPD   -5.933e-01  7.977e-02  -7.438 1.19e-11 ***
Sous_alimentation           -1.340e-02  1.546e-02  -0.867 0.387437
Acces_Internet_pers         -1.071e-02  9.971e-03  -1.074 0.284741
Acces_electricite_pers      -5.491e-04  8.579e-03  -0.064 0.949065
Pop_urban                   -2.550e-09  2.101e-09  -1.213 0.227224
Pop_rural                    3.252e-09  2.044e-09   1.591 0.113926
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.162 on 131 degrees of freedom
Multiple R-squared:  0.7492,    Adjusted R-squared:  0.7205
F-statistic: 26.09 on 15 and 131 DF,  p-value: < 2.2e-16
```

Nous avons ensuite accepté la normalité des résidus grâce au test de Shapiro-Wilk avec une *p-valeur=0.56.*

Nous avons comparé ce modèle avec le modèle constant mais le modèle constant est rejeté.

Nous avons fait une sélection de variables pour un sous-modèle avec le critère BIC et nous avons retenu les variables **Corruption, HDI et Budget militaire from GPD** pour expliquer la variable Gscore. Le test anova de sous-modèle est accepté avec p-valeur=0.37 et un R^2 de 0.702 ce qui semble plutôt raisonnable.

D'après le step AIC, le modèle qui explique le mieux le Gscore conserve les variables **Corruption**, **HDI**, **Budget militaire from GPD**, **Droits de l'Homme** et **Inégalité_des_genres**. On a un R^2 de 0,73 donc ça explique quand même très bien le Gscore. Et le sous-modèle est accepté par anova avec *p-valeur=0.73*

Finalement, nous pouvons penser qu'il n'est peut-être pas nécessaire de conserver autant de variables, car, sans trop pénaliser le R^2 on peut prendre en compte moins de variables ce qui simplifie le modèle.

Modèle Linéaire sur des pays très démocratiques et très autoritaires

Dans ce cas, nous avons sélectionné certains pays du jeu de données initial pour expliquer la variable Gscore. En fait, nous avons conservé uniquement les pays avec le gscore le plus élevé et ceux avec le gscore le plus faible pour avoir une distinction nette entre les différents pays. En effet, il est difficile de faire une réelle séparation entre un pays qui a un Gscore de 5.3 et un autre de 4 .8 et pourtant, ils sont classés dans des catégories différentes.

Ici, nous avons donc uniquement conservé les pays ayant un Gscore supérieur à 7.5 (les plus démocratiques) et ceux ayant un Gscore inférieur à 3.5 (les moins démocratiques).

Dans ce cas-là, on a expliqué le Gscore en fonction des mêmes variables que précédemment et nous avons un R^2 de 0.89 ce qui est très bien, les variables semblent très bien expliquer le niveau démocratique. De plus, les résidus suivent une loi normale.

Nous avons comparé ce modèle avec le modèle constant mais le modèle constant est rejeté.

Lorsque nous faisons appliquons la sélection de variables par critère **BIC**, nous conservons les variables **Droits de l'homme, Corruption et Budget militaire from GPD**. On retrouve presque les mêmes variables que lorsque nous avons travaillé avec tous les pays, seul **HDI** est remplacé par **Droits de l'Homme.** On a R^2=0.87 donc on ne perd que 0.02 par rapport au modèle de départ et il est accepté avec une *p-valeur=0.63.*

Avec le critère **AIC**, nous conservons les variables **Droits de l'Homme, Corruption, HDI, Budget militaire from GPD** et **l'Accès à Internet**. On a un R^2 de 0.88.

Il y a une très faible diminution de R^2 pour une importante diminution du nombre de variables étudiées. Il est donc peut être préférable de garder le sous modèle sélectionné avec la méthode bic pour ne pas avoir trop de variables, sans pour autant pénaliser la qualité du modèle.

Certaines variables sont toujours présentes quel que soit le modèle étudié. Nous pourrions donc penser que les variables qui font le plus de différence entre les pays démocratiques et les moins démocratiques sont : **Corruption, Budget militaire from GPD** et peut-être **Droits de l'Homme** (qui est présent dans 3 des 4 sous-modèles sélectionnés).

Modèle logistique

Nous avons ensuite fait un modèle logistique pour expliquer le Gscore. Pour cela, nous avons transformé la variable quantitative en variable binaire : Lorsque le Gscore était supérieur à 5, nous avons considéré le pays démocratique et nous lui avons donné la valeur 1. Lorsque le Gscore était inférieur à 5, nous avons considéré le pays non démocratique et lui avons attribué une valeur égale à 0. Nous avons donc utilisé *glm* avec la fonction lien *« logit ».*

Premièrement, nous avons expliqué le Gscore en fonction de toutes les autres variables et nous avons obtenu un pseudo R^2 de 0.5 ce qui est moins bien que pour les modèles linéaires.

Grâce aux odds ratios, nous avons pu remarquer que :

- Si la variable Droits de l'Homme augmente de 1 unité, alors la chance que le pays soit démocratique est multipliée par 1.5 (cad il y a plus de chance que le pays soit démocratique).

Nous avons ensuite fait une procédure de sélection de variables avec le critère AIC, et nous avons conservé les variables : **Corruption, Annees_detudes, Espérance_vie, Budget_militaire_from_GPD et Acces à Internet,** avec un pseudo R^2 de 0.46 ce qui n'est pas trop mal au vu du nombre de variables supprimées par rapport au modèle initial. Et ce sous-modèle est accepté par anova avec une *p-valeur=0.83.*

De plus avec le critère BIC, on a gardé les variables : **Corruption, Années d'études et Budget militaire from GPD** et on a R^2 de 0.44. Ce sous-modèle est accepté par anova.

⇨ **De plus, on retrouve une partie des variables que nous avions avec le modèle linéaire complet**

Modèle logistique sur des pays très démocratiques et très autoritaires

Nous avons ici créé un sous jeu de données contenant uniquement les pays ayant un Gscore élevé >7.5) et un Gscore faible (<3.5) et nous avons transformé cette même variable en variable binaire pour mettre en place un modèle logistique.

Pour le modèle logistique complet nous avons obtenu un R^2=0.8. La valeur étant très proche de 1, nous pouvons dire que les variables expliquent bien le Gscore.

```
Call:
glm(formula = Gscore ~ ., data = data3)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.57916  -0.08980  -0.00441   0.09644   0.56196

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -9.600e-01  8.275e-01  -1.160   0.2515
Droit_de_lHomme               4.769e-02  3.520e-02   1.355   0.1816
Liberte_economique            3.883e-02  5.776e-02   0.672   0.5046
Inegalite_des_genres         -2.880e-01  3.704e-01  -0.778   0.4405
Proportion_denfants_harceles -1.519e-03  3.078e-03  -0.494   0.6237
Corruption                    8.167e-03  4.093e-03   1.996   0.0515 .
Taux_alphabetisation         -1.935e-03  3.067e-03  -0.631   0.5309
HDI                           3.744e-01  1.005e+00   0.372   0.7112
Annees_detudes               -2.287e-03  2.809e-02  -0.081   0.9355
Espérance_vie                 1.575e-02  1.339e-02   1.176   0.2451
Budget_militaire_from_GPD    -6.108e-02  2.361e-02  -2.588   0.0126 *
Sous_alimentation             2.985e-03  5.388e-03   0.554   0.5821
Acces_Internet_pers          -3.460e-03  2.938e-03  -1.177   0.2446
Acces_electricite_pers       -9.814e-04  2.635e-03  -0.372   0.7112
Pop_urban                     1.991e-09  1.046e-09   1.904   0.0626 .
Pop_rural                    -3.827e-09  1.758e-09  -2.177   0.0343 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.05256145)

    Null deviance: 16.2576  on 65  degrees of freedom
Residual deviance:  2.6281  on 50  degrees of freedom
AIC: 8.5552

Number of Fisher Scoring iterations: 2
```

```
```{r}
pseudoR2 = (glm.Gscore2$null.deviance - glm.Gscore2$deviance)/glm.Gscore2$null.deviance
pseudoR2
```
```

```
[1] 0.8383478
```

Le modèle constant a été rejeté avec une p-valeur inférieur à2.2e-6.

Le critère AIC nous propose de conserver **Droits de l’Homme, Corruption, Espérance de vie, Budget militaire from GPD, Accès à Internet, Population rurale et urbaine**. Afin d’étudier ce modèle nous avons mis des paramètres de contrôle dans la fonction *glm* soit *maxit=100* et *epsilon=1* car sans ces paramètres l’algorithme ne convergeait pas. On a alors R^2=0.74 mais le sous-modèle est rejeté par anova.

En revanche avec le critère AIC, nous conservons les variables **Corruption, Budget militaire from GPD, Pop urbaine et Pop rurale** avec un R^2=0.85 qui est plus élevé que celui du modèle complet**. Cependant ce sous modèle est rejeté par anova.**

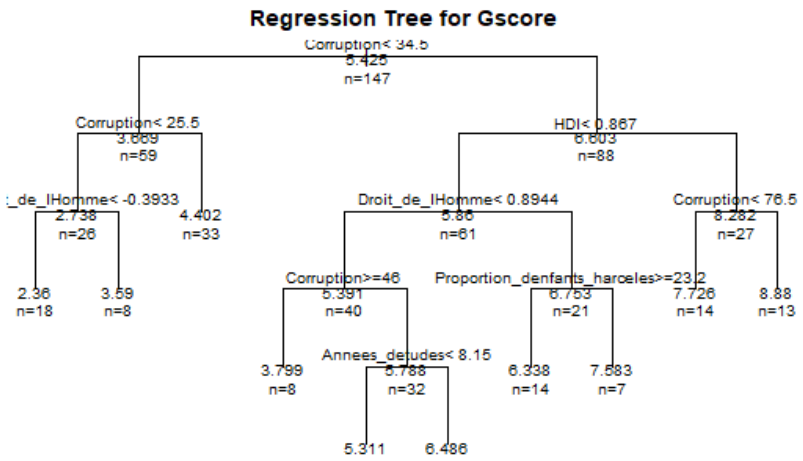
Tableau résumant la situation :

	Mod lin complet	Mod lin incomplet	Mod logit complet	Mod logit incomplet
AIC	Corruption HDI Budget militaire from GPD Droits de l'Homme Ingalité_des_genres	Droits de l'Homme Corruption HDI Budget militaire from GPD Accès à Internet	Corruption Annees_detudes Espérance_vie Budget_militaire_from_GPD Acces à Internet	Modèle rejeté
BIC	Corruption HDI Budget militaire_from GPD	Droits de l'homme Corruption Budget militaire from GPD	Corruption Années d'études Budget militaire from GPD	Modèle rejeté

Arbre de Régression et Classification

Nous avons tout d'abord fait des arbres de régressions et de classification sur l'ensemble du jeu de données pour sélectionner les variables qui permettent de faire la meilleure distinction entre les pays démocratiques et les pays non démocratiques.

Arbre de régression sur le jeu de données complet

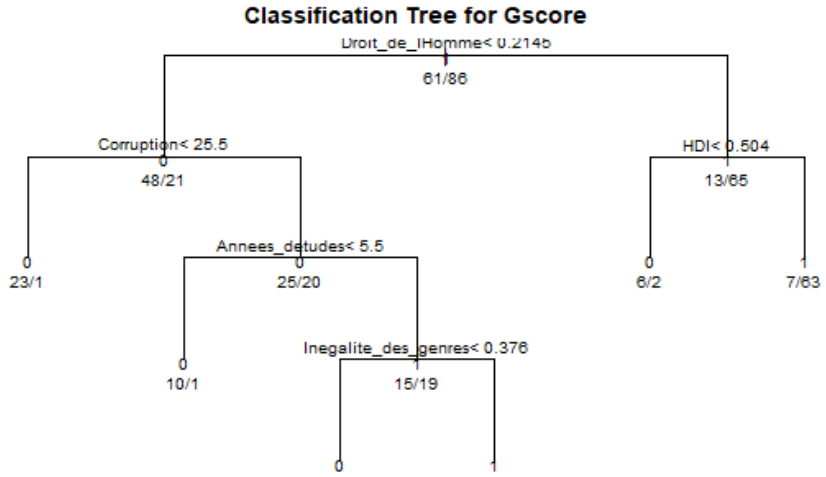


Il semblerait que les variables qui permettent de distinguer les pays les plus démocratiques sont la variable **Corruption** avec un « taux » inférieur à 34.5 ainsi qu'un **HDI** <0.867 puis de nouveau **Corruption** <76.5. Il y a 13 pays dans le jeu de données qui respectent ces trois conditions, avec un score démocratique moyen de 8.88.

En revanche, pour déterminer les pays les moins démocratiques, il semble que les critères les plus importants à prendre en compte sont : **Corruption** <25.5 et **Droits de l'Homme**<-.03933 (un taux négatif signifiant que les droits de l'Homme sont peu respectés) .

Arbre de classification sur le jeu de données complet

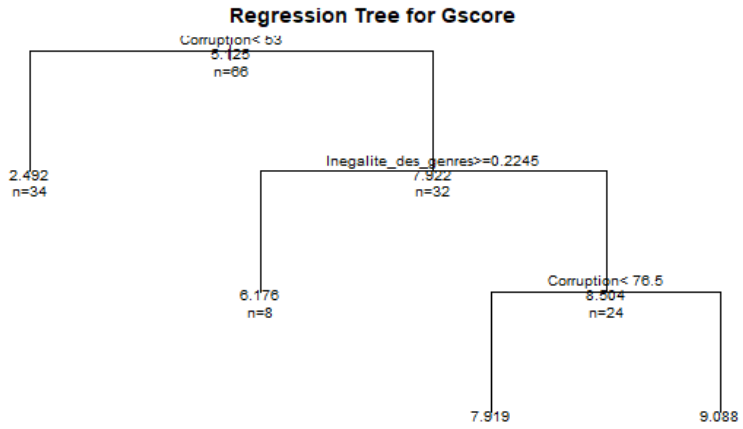
Sans paramètre



Ici, nous avons transformé le Gscore en variable binaire, et les variables les plus importantes semblent être **Droits de l'Homme, Corruption, HDI, années d'études ett Inégalité des genres**. Malgré quelques différences, nous avons des variables en commun avec les arbres de régression. La seul différence est **Inégalités des genres** en classification et **Proportion d'enfants harcelés** en Regression.

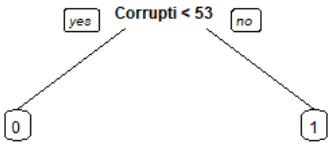
Nous allons maintenant mettre en place ces arbres sur un sous jeu de données, en conservant uniquement les pays les plus démocratiques et les pays les moins démocratiques.

Arbre de régression sur les pays très démocratiques et très autoritaires



Seul les variables **Corruption** et **Inégalités des genres** ont été gardées.

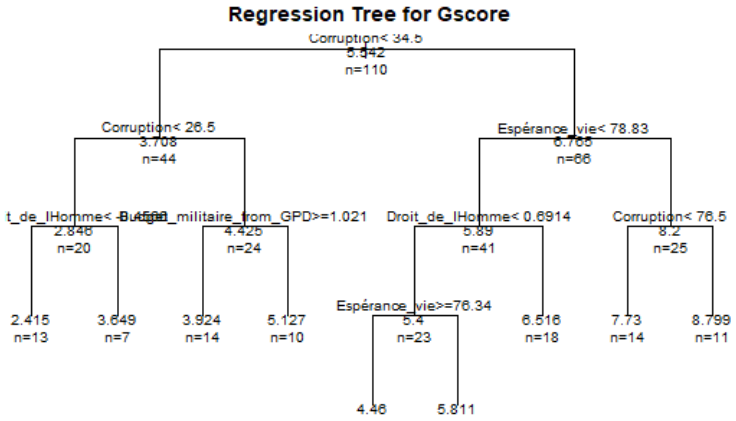
Arbre de classification sur un sous jeu de données



Le nombre de variables ayant été réduit à 2 en régression et à 1 en classification, nous ne sommes pas sûrs de la pertinence des résultats. Cependant, les 2 arbres présentés précédemment permettent de souligner à nouveau l'importance de la variable **Corruption**.

Arbre de régression et classification pour la prédiction

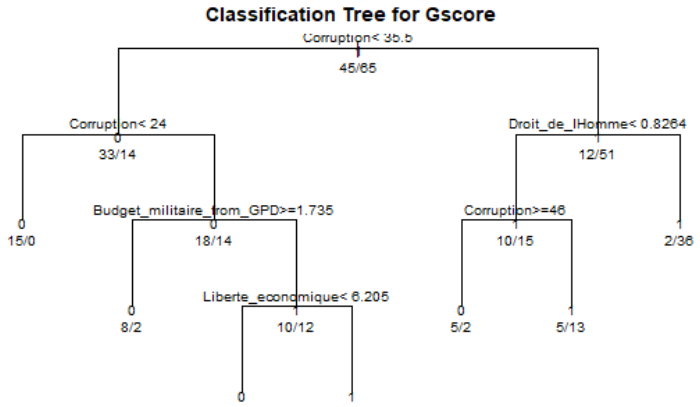
Nous avons enfin mis en place des arbres de régression afin de savoir s'il permettait de prédire convenablement le score démocratique du pays.
Pour cela, nous avons décomposer le jeu de données initial en 2 échantillons :
- un **échantillon d'apprentissage** contenant 75% des pays
- un **échantillon test** contenant 25% des pays



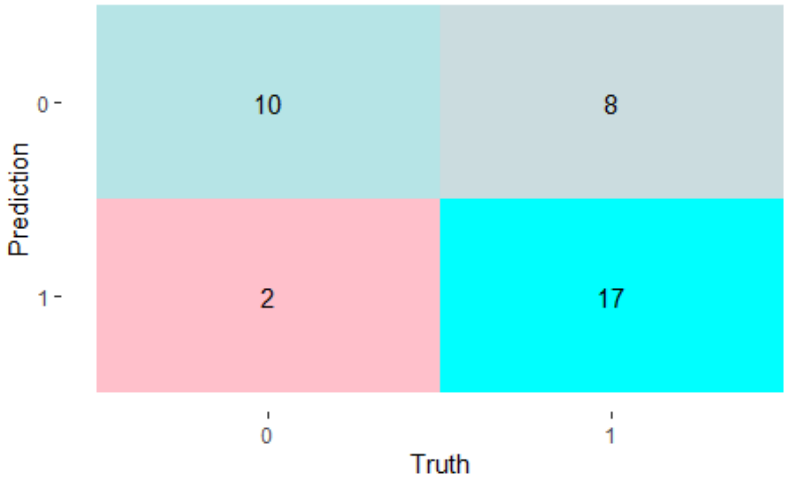
Sachant que l'échantillon d'apprentissage est aléatoire, nous n'avons pas toujours le même arbre de prédiction lorsque à chaque fois que l'algorithme est exécuté, mais des variables semblent revenir à

plusieurs reprises tel que : **Corruption, Budget militaire from GPD, Droits de lHomme** et **Accès_Internet**. Cependant, la qualité de la prédiction n'est pas très bonne. En effet, de nombreux pays ont un score prédits très différents de plus de 3 points de leur Gscore selon The Economist.

La qualité de la prédiction semble légèrement meilleure lorsque nous mettons en place des arbres de classification. Ici aussi le résultat varie sachant que l'échantillon d'apprentissage est aléatoire. Les variables qui reviennent le plus souvent sont **Corruption** et **Droits de l'Homme**.



Voici un exemple d'arbre de classification et les résultats de prédiction obtenus :



Il y a plus d'erreur sur des pays qui sont prédits démocratiques alors qu'ils sont autoritaires que l'inverse. En effet, si on regarde le Gscore on voit que souvent les pays avec un Gscore entre 4,5 et 5 sont prédits comme démocratiques. Plus quelques exceptions.

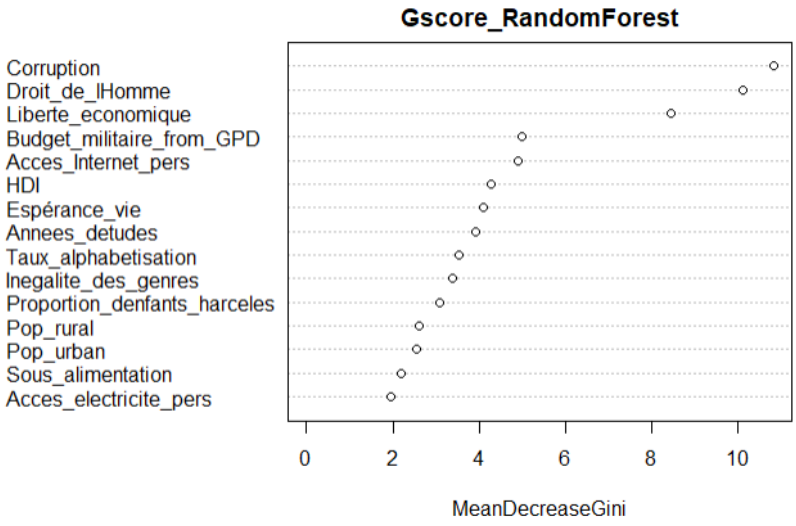
Random Forest

Jeu de données complet

Nous avons enfin utilisé les random forest pour mettre en avant les variables les plus pertinentes afin d'expliquer le niveau démocratique.

Dans le cas de la **régression**, les cinq variables qui semblent le plus à même d'expliquer le Gscore sont **Corruption, Espérance de vie, Droits de l'Homme et HDI**.

Dans le cas de la **classification**, nous avons pu remarquer que les variables **Corruption, Droits de l'Homme et Liberté économique** semblent le mieux expliquer le régime démocratique.



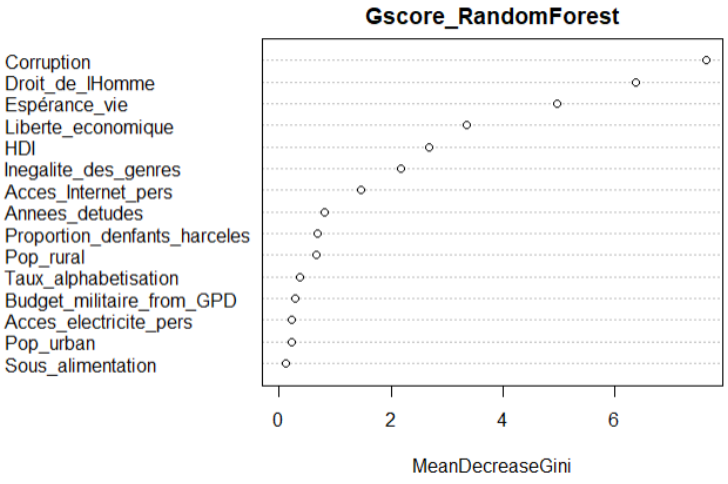
Il y adonc les varaibles **Corruption** et **Droits de l'Homme** qui sont communes.

Sur les pays très démocratiques et très autoritaires

Nous conservons (comme dans les analyses précédentes) uniquement les pays les plus démocratiques et uniquement les pays les moins démocratiques.

- Dans le cas de la **régression**, nous conservons les variables **Droits de l'Homme, Corruption et Espérance de vie**. C'est très proche des variables trouvées avec le jeu de données complet.

- Dans le cas de la **classification**, c'est **Corruption, Droits de l'Homme et Espérance de vie**, soit les même qu'en régression.

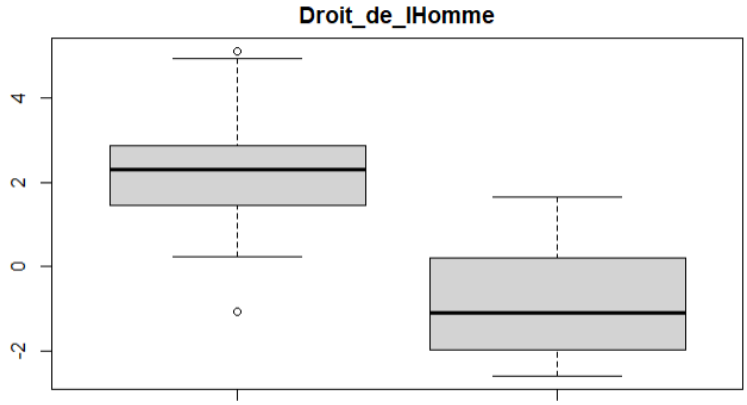


Ressemblance entre les pays démocratiques et non démocratiques

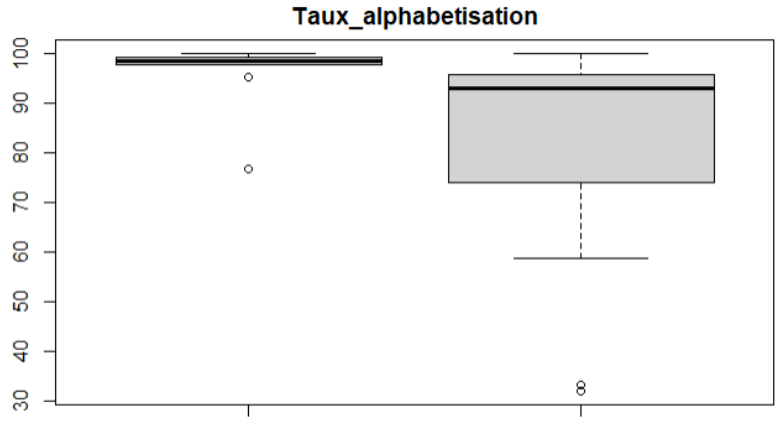
Ici, à partir du jeu de données initial, nous avons construit 2 sous jeux de données.

- Data1= qui contient les pays ayant un Gscore supérieur à 7.5
- Data2= qui contient les pays ayant un Gscore inférieur à 3.5

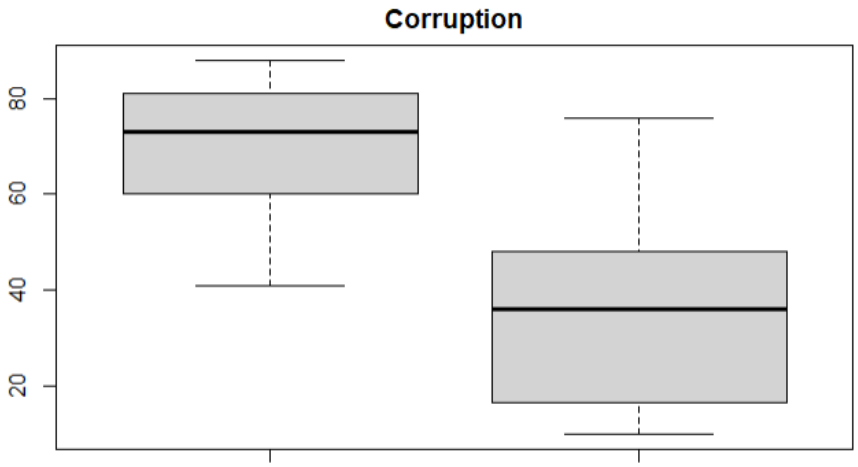
Nous avons pour chaque variable de chaque jeu de données représentés des boxplots pour voir la différence d'étendu des variables. Nous avons également fait des tests de comparaisons de moyennes (*test de Wilcoxon*) et de variance (*test d'ansari*) pour savoir si elles étaient bien différentes pour ces 2 catégories de pays. Plus précisément le test d'ansari teste si les échantillons diffèrent uniquement par leur variance.



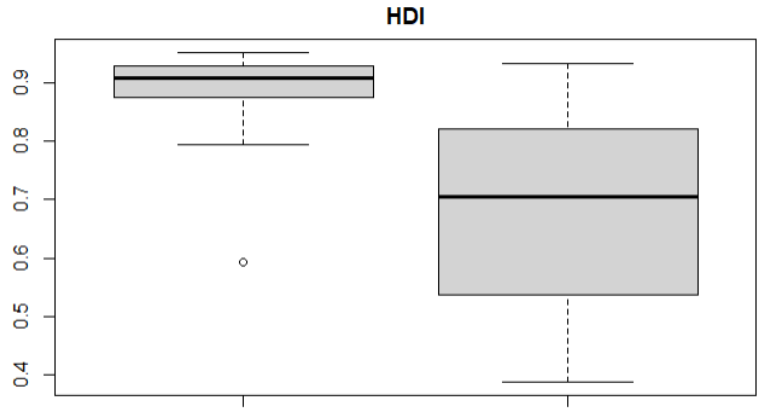
Par exemple, dans le cas des Droits de l'Homme, on peut clairement voir que les pays les plus démocratiques ont un indice positif à l'exception de l'outlier tandis que ¾ des régimes autoritaires auront tendances à être négatifs. En revanche, il n'y a pas de différence significative en termes de dispersion des données.



Pour le taux d'alphabétisation, les moyennes sont assez proches mais la dispersion pour les pays non démocratiques est nettement plus importante.



Les boxplots pour les deux échantillons sont plutôt indépendants avec une médiane haute pour les pays démocratiques qui signifie que les pays est peu corrompu.



On remarque une grande dispersion dans les données pour les autocraties contrairement aux données des démocraties qui sont plutôt proches à l'exception d'un outlier.

Conclusion :

Nous pouvons remarquer que globalement, les pays démocratiques semblent avoir des variances et des dispersions plus faibles pour les variables considérées par rapport aux pays démocratiques. Cependant, avec cette étude il semble difficile de dire que les pays démocratiques sont tous "identiques".

Il semblerait que les pays démocratiques soient plus ou moins similaires pour les critères :

- Droit de l'homme avec un indice positif

- Liberté économique avec une faible dispersion avec 3/4 des valeurs entre 7.4 et 7.5

- Inégalité des genres avec 3/4 des valeurs entre 0.06 et 0.13

- Corruption : 1er quartil à 71 donc 3/4 des valeurs au-dessus de 71 tandis que dans les régimes autoritaires ont 3/4 des valeurs sont en dessous de 48

- Taux alphabétique avec 3/4 des valeurs entre 97 et 99

- HDI avec 3/4 des valeurs entre 0.87 et 0.93

- Espérance de vie avec 3/4 des valeurs entre 81 et 83 ans

- Budget militaire avec 3/4 des valeurs entre 0.9 et 1.8 (et avec le 1er quartile des régimes autoritaires à 1.8)

- Sous-alimentation tous à 2.5 sauf 5 outliers

Mais ce n'est pas le cas pour les critères comme :

- Electricité car la dispersion est faible mais trop proche des régimes autoritaires.

- Pour la variable proportion d'enfants harcelés les deux échantillons ont des moyennes très proches et il n'y pas grande différence en termes de dispersions

Quant aux pays non démocratiques, il y a très peu de critères qui ressortent (droit de l'homme entièrement négatif, corruption avec 3/4 des valeurs en dessous de 48 et un budget militaire avec 50% des valeurs entre 1.8 et 4.1)

Il ne semblerait pas qu'on puisse conclure qu'ils sont tous identiques puisque nous avons aussi des dispersions globalement importantes pour la majorité des variables.

On pourrait donc dire qu'il y a plusieurs "façons" d'être démocratique (même si la richesse du pays semble un critère important) et plusieurs façons de ne pas l'être.

Enfin avec le test de Shipiro-wilk nous avons relevé les échantillons des variables **Droits de l'Homme et Liberté économique** suivaient une loi normale

Gestion des données manquantes

Pour certains cas, des données étaient manquantes, nous avons utilisé pour certaines analyses un algorithme de miss forest afin de compléter le jeu de données.

CONCLUSION

En conclusion, même si nous retrouvons des résultats un peu différents avec nos nombreuses analyses certaines variables reviennent souvent et sembleraient nous permettre d'expliquer le niveau démocratique d'un pays. En effet, il est possible de faire le lien entre le Gscore de the Economist et des variables tel que la Corruption, Droits de l'Homme, Liberté économique et Budget militaire from GPD. Il persiste toujours des indices liés à l'économie du pays mais pas uniquement