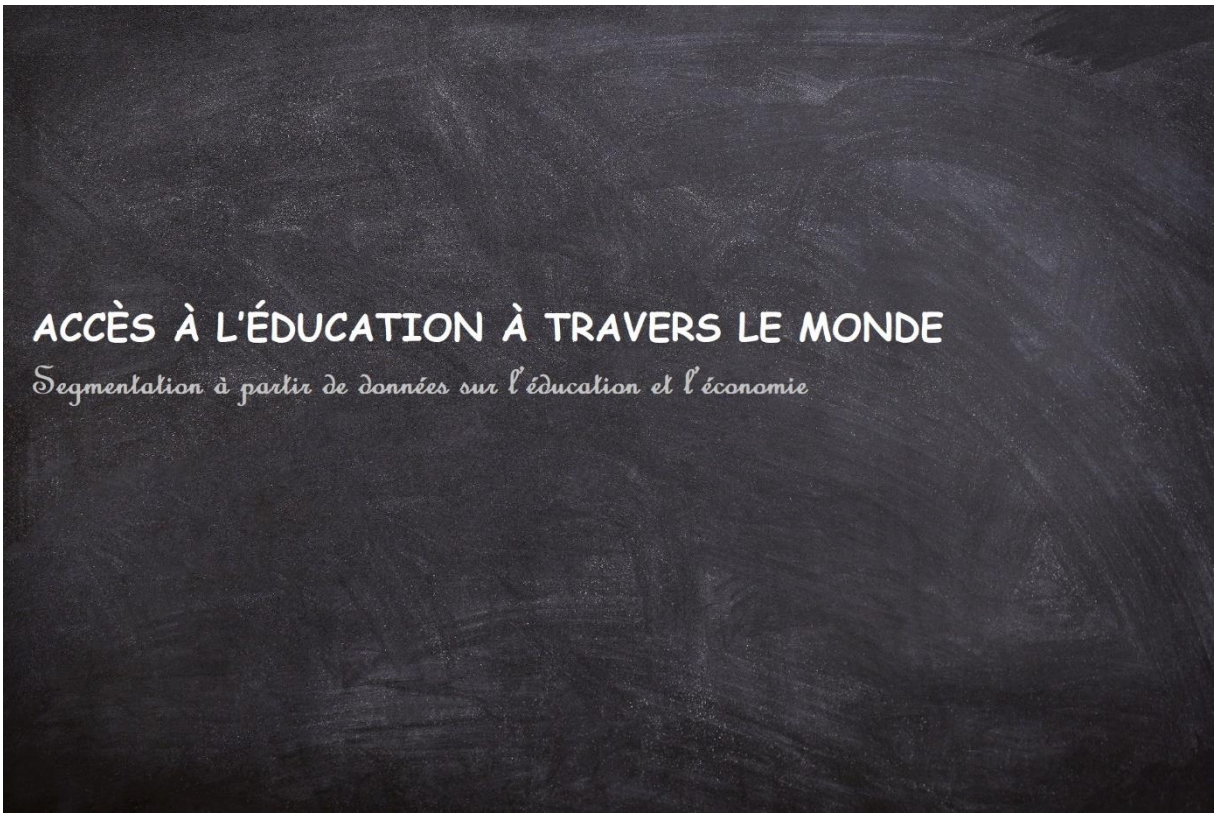


Cécile Guillot – Data Analyst

Table des matières

Index des tableaux

Tables des figures

Introduction

L'accès à l'éducation est une chose essentielle pour chaque enfant. Il permet d'acquérir des connaissances pour en faire des citoyens avisés ou encore de rencontrer d'autres enfants d'horizons différents. Cependant, on remarque qu'à travers le monde, l'éducation n'est pas accessible à tous.

Pour améliorer l'accès à l'éducation pour tous, il peut être intéressant d'observer les caractéristiques des pays où cet accès est facilité mais aussi les caractéristiques des pays où il y a de forts taux de non-scolarisation.

Tout d'abord, nous allons donc faire un état des lieux de l'accès à l'éducation à travers le monde. Ensuite, nous nous intéresserons à la santé économique de chacun de nos pays. Enfin, à partir de ces données, nous allons créer des groupes en utilisant des méthodes de classifications et des algorithmes non-supervisés. Plus spécifiquement, ce projet se base sur des techniques de segmentations utilisées dans le marketing pour créer des groupes de clients. Plusieurs algorithmes seront testés pour voir celui qui permet d'obtenir la classification la plus fine possible. Une fois cette étape réalisée, nous réaliserons une description des groupes obtenus pour élaborer des pistes d'amélioration pour les pays où l'accès à l'éducation est le plus faible.

Ce projet s'adresse donc à des organisations telles que l'UNICEF ou l'UNESCO qui s'intéressent à la scolarisation de tous à travers le monde. Il peut aussi représenter un intérêt pour les gouvernements des pays pour avoir une idée des actions à mettre en place pour améliorer l'accès à l'éducation dans leur pays.

Objectifs :

- Dresser un portrait de la situation actuelle concernant l'accès à l'éducation à travers le monde ;
- Segmenter nos pays sur la base de données portant sur l'éducation et la santé économique ;
- Décrire les caractéristiques de nos groupes pour pouvoir avoir des bases sur les actions à mettre en place.

Acquisition des données

Les données utilisées dans cette étude ont été téléchargés à partir de la base de données de la banque mondiale (cf Annexe 2). Tous les pays ont été sélectionnés mais certains ont été supprimés car ils ne représentaient pas d'intérêt. Au total, une trentaine de variables ont été sélectionnés pour réaliser une description de l'accès à l'éducation et la santé économique à travers le monde.

Etat des lieux de l'éducation à travers le monde

L'état des lieux de l'éducation à travers le monde a été réalisé sur 204 pays. Les données exploitées ont été téléchargés sur le site de la banque mondiale. Les enquêtes sur la scolarité n'étant pas annuelles, il a été fait le choix de prendre les chiffres des dix dernières années et de faire une moyenne pour chaque pays et chaque variable. Ainsi, cela nous permet d'obtenir un nombre de données assez importants. Les variables qui ont servi à décrire cette situation sont au nombre de 17 et portent sur le taux de non-scolarisation, les inscriptions ainsi que le ratio de genre pour chaque cycle scolaire.

Les moyennes mondiales

Les moyennes de nos différentes variables sont présentées dans le tableau ci-dessous. Les variables ont été séparés par type pour mieux faciliter leur lecture.

Enfants non-scolarisés	8.48 %
Adolescents non-scolarisés	14.4 %
Inscriptions en préscolaire	58.2 %
Inscriptions en primaire	103 %
Inscriptions en cycle secondaire	77.6 %
Inscriptions en enseignement supérieur	34 %
Taux d'achèvement du cycle primaire	88.3 %
Taux d'achèvement du cycle secondaire	73.6 %

Ratio filles/garçons dans l'enseignement primaire	0.96
Ratio filles/garçons dans l'enseignement secondaire	0.96
Ratio femmes/hommes dans l'enseignement supérieur	1.05
Taux d'alphabétisation des jeunes	89 %
Taux d'alphabétisation total	82 %

Tableau I. Moyenne des différentes variables utilisées pour faire l'état des lieux de l'éducation à travers le monde.

L'âge moyen de début du cycle primaire est de 6 ans et l'âge moyen de début du cycle secondaire est de 11.8 ans.

On observe que seulement 8.5 % des enfants et 14.5 % des adolescents en âge de l'être ne sont pas scolarisés. Ces chiffres peuvent sembler peu élevés pour des cycles qui sont obligatoires pour tous mais la moyenne ne traduit pas les disparités que l'on peut trouver à travers le monde.

Des données disparates

A travers le monde, on estime que la population entre 0 et 14 ans correspond à presque 30 % de la population mondiale. Selon les informations ci-dessus, cela correspond à des enfants pouvant fréquenter le cycle primaire et des adolescents pouvant fréquenter le cycle secondaire. Pour mieux comprendre en quoi nos données sont disparates, la carte ci-dessous présente le pourcentage d'enfants et d'adolescents non-scolarisés à travers le monde.

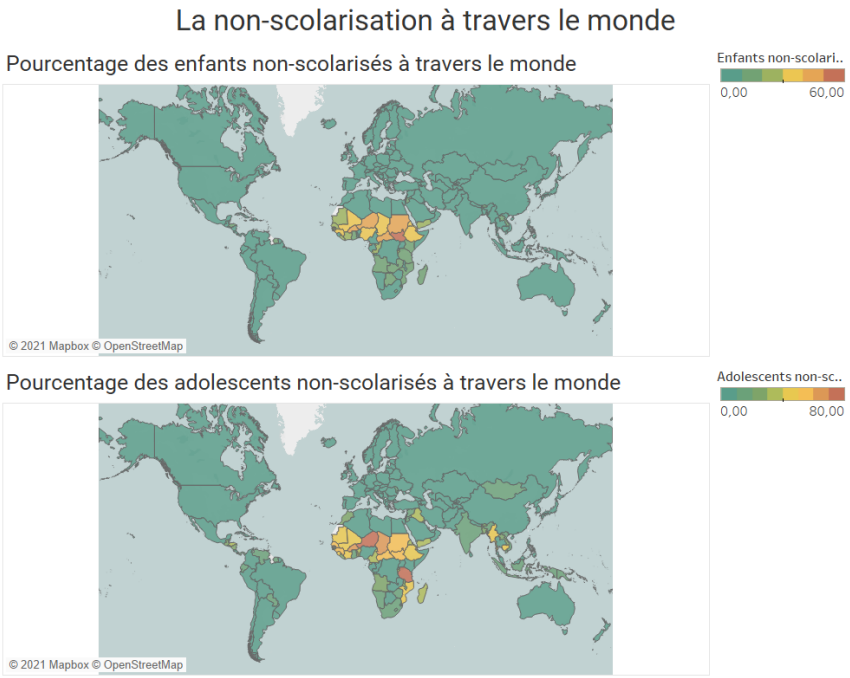


Figure 1. Carte du pourcentage de non-scolarisation à travers le monde pour les enfants et les adolescents

Les pays d'Afrique centrale sont ceux où les taux de non-scolarité sont les plus élevés. On remarque d'ailleurs que lorsque le taux est important pour les enfants en âge d'intégrer le cycle primaire, ce taux devient d'autant plus important que les enfants atteignent l'adolescence.

Lien entre les variables

Les variables sélectionnées sont de plusieurs types. Elles vont s'intéresser aux taux d'inscriptions dans les différents cycles, aux taux de non-scolarisation, aux taux d'achèvement des différents cycles ou encore à l'alphabétisation des jeunes et total.

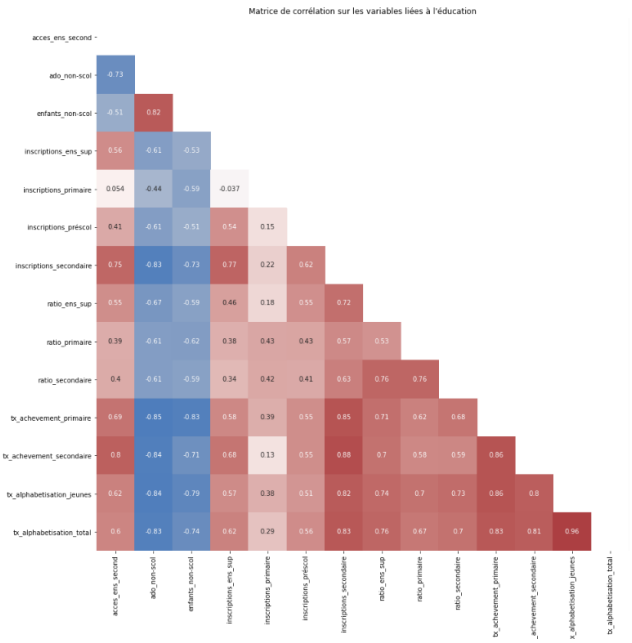


Figure 2. Matrice de corrélation des variables en lien avec l'éducation

Il est intéressant de voir que le taux d'enfants non-scolarisés semble être un facteur prédictif du taux d'adolescents non-scolarisés. De plus, le taux d'achèvement du primaire va être prédictif du taux d'inscriptions en cycle secondaire.

Plus simplement, on remarque que les données du cycle primaire semblent être les plus déterminantes. En effet, le cycle primaire est présent et/ou obligatoire dans de nombreux pays contrairement aux cycles préscolaires. Ainsi si les enfants sont inscrits dès leur 6 ans à l'école, ils ont plus de chance d'aller jusqu'à la fin du cycle secondaire.

Les ratios sont équilibrés dans les cycles scolaires obligatoires mais les disparités apparaissent lors des inscriptions en enseignement supérieur.

Conclusion

Grâce à nos données sur l'éducation, on voit déjà apparaître des inégalités concernant l'accès à l'éducation. Les pays d'Afrique centrale, mais aussi de la péninsule arabique, montre de fort taux de non-scolarisation des jeunes. Pour comprendre cette tendance, il peut être intéressant de se pencher sur la santé économique des pays à travers le monde. En effet, on pourra ainsi

voir s'il y a des indices économiques qui peuvent permettre d'établir un lien avec nos données éducatives.

La santé économique à travers le monde

Comme pour les données sur l'éducation, les données sur la santé économique ont été téléchargées à partir de la base de données de la banque mondiale. Pour un minimum de cohérence, les chiffres des 20 dernières années ont été téléchargés et ce sont les moyennes des 20 dernières années qui nous ont permis de construire notre jeu de données. Nous avons donc accès à 12 variables qui ont été collectés parmi 217 pays.

La santé économique mondiale

Les données présentes dans le tableau ci-dessous sont les moyennes calculées pour les 217 pays de notre étude.

Capacité/Besoin de financement	0.56 %
Croissance du PIB	3.57 %
Croissance du PIB par habitant	2.13 %
Epargne brute (% du PIB)	23 %
Epargne intérieure brute (% du PIB)	18.1 %
Indice GINI	43.18
Part des revenus détenus par les 20% moins élevés	6.48 %
Part des revenus détenus par les 20% plus élevés	46 %
Population active avec un niveau d'étude de base	51.8 %
Chômage	8.6 %
Taux d'emploi des 15 ans et plus	56.10 %
Taux d'activité des 15-24 ans	43.6 %

Tableau II. Moyenne des différentes variables utilisées pour faire l'état des lieux de l'économie dans le monde.

En dépit de moyennes mondiales plutôt correctes, on remarque là aussi que les données sont disparates à travers le monde. On va s'intéresser plus particulièrement à l'indice de GINI.

La répartition des richesses à travers le monde

L'indice de GINI est un indice économique marqueur de la répartition de la richesse à travers un pays. Sa sélection le rend plus pertinent qu'une croissance annuelle du PIB par habitant. En effet, il existe des pays où la croissance du PIB par habitant va être importante mais où les richesses seront mal réparties.

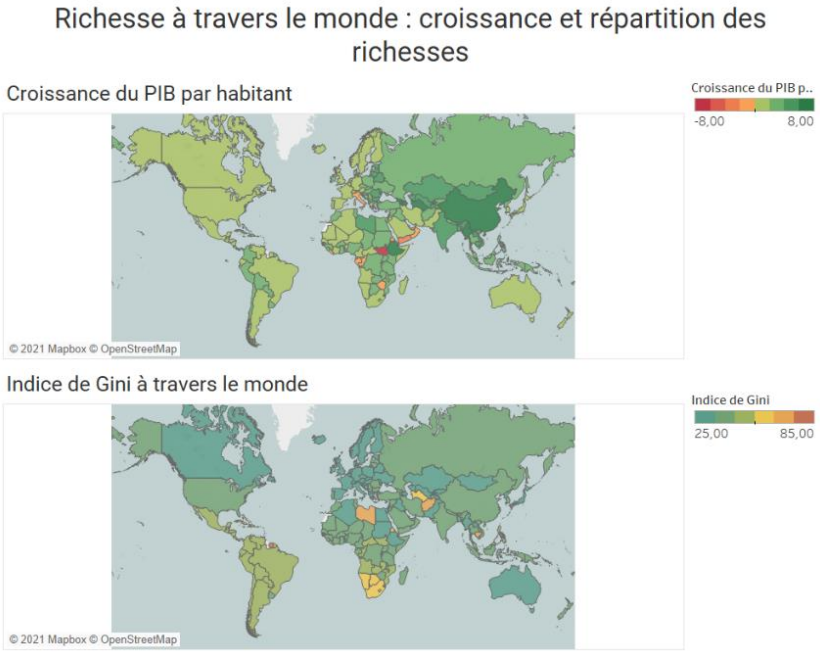


Figure 3. Carte du PIB par habitant et de l'indice de GINI dans les pays du monde

Tout d'abord, on remarque que selon l'indice de GINI, il y quelques pays où les richesses sont mal réparties. L'Afrique du Sud ou encore la Lybie viennent appuyer le propos selon lequel les pays avec une forte croissance du PIB par habitant peuvent avoir une répartition inégale des richesses.

Cependant, on remarque que les pays d'Afrique où les taux de scolarité sont les plus faibles ne sont pas tous ceux avec une croissance du PIB par habitant faible et/ou une mauvaise répartition des richesses.

La création de notre modèle devient d'autant plus intéressante que l'on peut s'attendre à voir émerger des groupes où la scolarisation est un problème sans que cela ne soit dû à un manque de richesses.

Néanmoins, avant de passer à la modélisation, il est intéressant de regarder s'il y a des corrélations entre nos différentes variables économiques.

Lien entre les variables sur la santé économique mondiale

Comme pour les variables sur la scolarité, on a cherché à voir s'il y avait des corrélations qui pouvaient nous aider à mieux décrypter la santé économique mondiale.

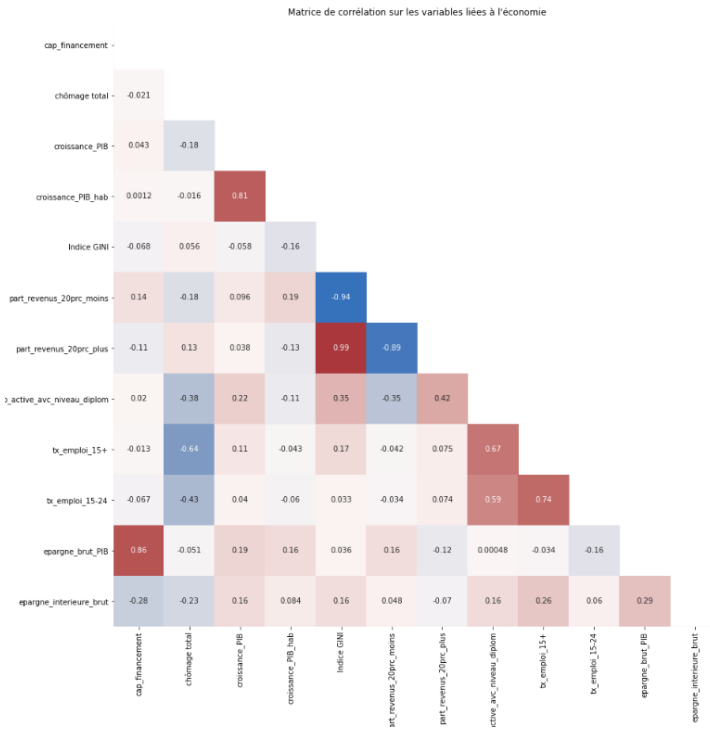


Figure 4. Matrice de corrélation des variables en lien avec l'économie

On remarque que les capacités de financement d'un pays sont corrélées à l'épargne brut. Ensuite, les variables en lien avec la répartition des richesses sont corrélées négativement avec

l'indice de Gini. Enfin, on observe que la possession d'un diplôme va avoir un impact sur l'employabilité des 15 ans et plus.

Conclusion

L'analyse des données en lien avec la santé économique ne montre pas autant d'évidence que les données sur l'éducation. En effet, la distinction entre pays riches et pays pauvres semble être quelque chose de moins à voir. De plus, il semblerait que ce facteur ne soit pas décisif dans la non-scolarisation des enfants.

L'utilisation d'algorithme de classification permettra donc de créer des groupes de pays avec des particularités proches. Ainsi la mise en place d'action pour améliorer la scolarisation des enfants et des adolescents sera plus adaptée.

Modélisation : Utilisation d'algorithmes de clustering

Les techniques de machine learning peuvent permettre de créer des groupes pour mieux mettre en place des plans d'action pour permettre à tous les enfants et adolescents d'accéder à l'éducation. Pour répondre à notre problème, les algorithmes de classification non-supervisée vont nous être utile. Dans la suite de cette étude, nous allons faire appel à l'algorithme du K-Means, la classification ascendante hiérarchique et l'algorithme DBSCAN. A l'issue du développement de ces algorithmes, on pourra comparer les groupes obtenus pour avoir une idée de celui qui semble le plus performant.

L'algorithme du K-Means

L'algorithme du K-Means permet de créer des clusters en spécifiant le nombre de clusters que l'on souhaite obtenir. Pour déterminer le nombre de clusters, on utilise la méthode du coude qui se base sur les distorsions des variables de notre étude. Cette méthode nous montre que le nombre idéal de clusters serait de 4 ou 5.

K = 2

Avant de modéliser un algorithme de K-Means avec K = 4 ou K = 5, on va effectuer un test avec K = 2 pour voir si on peut voir ressortir les pays avec des taux de non-scolarisation élevée.

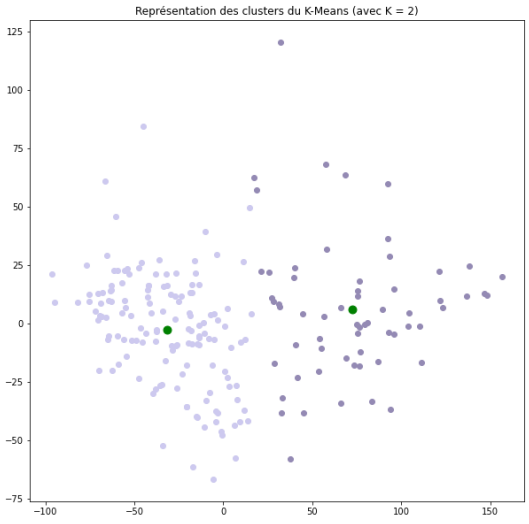


Figure 5. Représentation des clusters de l'algorithme du K-Means avec K = 2

L'algorithme du K-Means nous a permis de déterminer deux clusters que l'on a projeté dans les deux premières de projections d'une analyse en composantes principales. Pour rendre cela plus lisible, on va projeter nos clusters sur une carte du monde. Cela sera présenté dans la partie « Analyse de nos modèles ».

K = 4

La méthode du coude nous a permis de voir que le nombre optimal de clusters est de 4 ou 5. On va d'abord commencer par entraîner un algorithme avec K = 4.

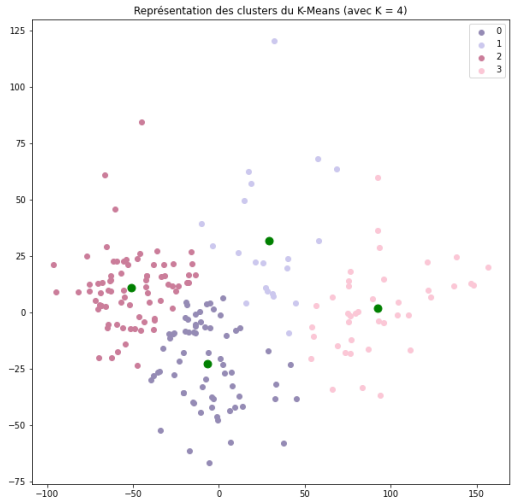


Figure 6. Représentation des clusters de l'algorithme du K-Means avec K = 4

L'algorithme du K-Means avec K = 4 nous permet d'obtenir quatre groupes plutôt bien équilibrés et bien répartis sur les deux premières composantes de l'ACP. Les caractéristiques de nos quatre groupes seront détaillées dans la suite de ce rapport.

K = 5

La méthode du coude nous montrait des résultats intéressants pour un K = 4 ou 5. On a donc entraîné un algorithme du K-Means avec K = 5. Les clusters sont représentés dans les deux premières composantes de l'ACP.

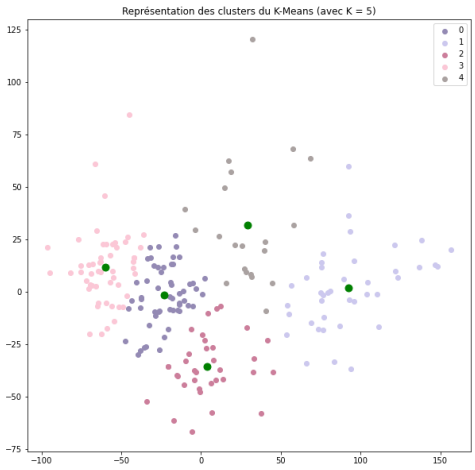


Figure 7. Représentation des clusters de l'algorithme du K-Means avec K = 5

L'algorithme du K-Means avec K = 5 a permis de créer 5 clusters avec un nombre de pays plutôt équilibré. Le détail de chaque groupe sera vu dans la dernière partie de son rapport.

La classification ascendante hiérarchique : le dendrogramme

Un deuxième type d'algorithme qui peut être utiliser pour effectuer des classifications est la classification ascendante hiérarchique. Ici, on a effectué une classification hiérarchique en définissant 3 clusters. Les résultats sont présentés ci-dessous.

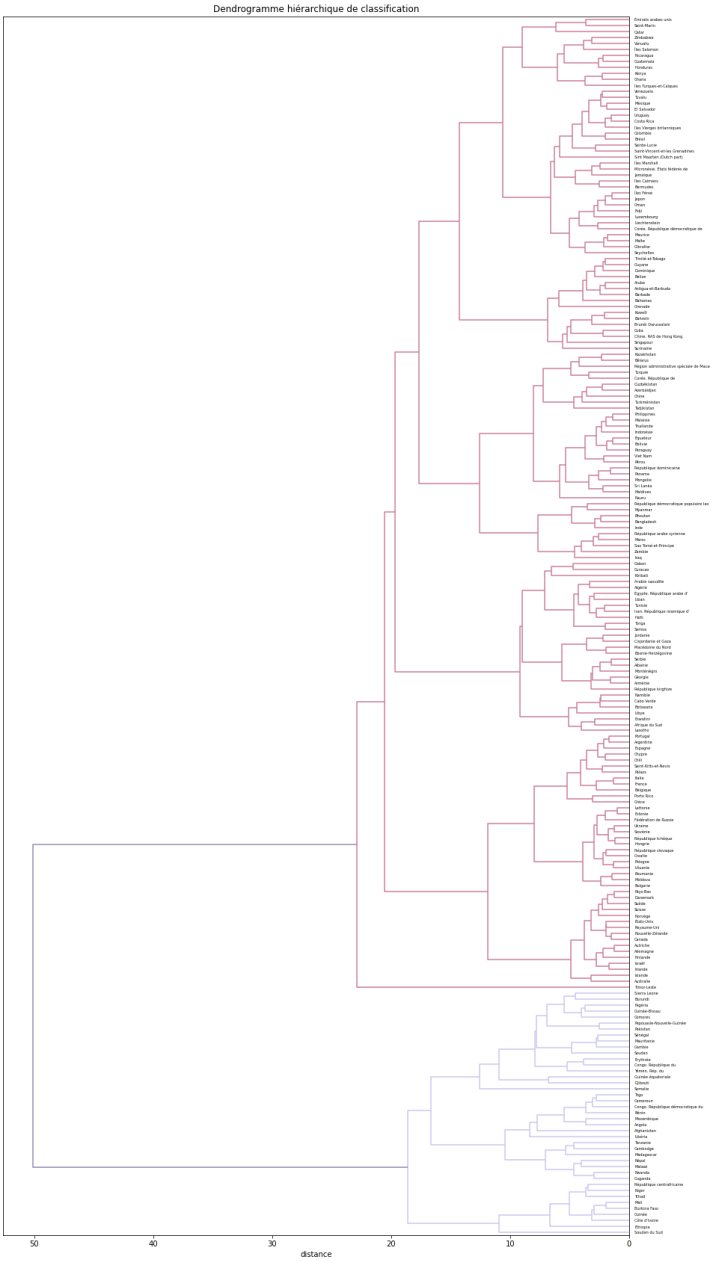


Figure 8. Dendrogramme de la classification ascendante hiérarchique des pays avec 3 clusters

L'algorithme du DBSCAN

Enfin, le dernier type d'algorithme de classification utilisé dans ce projet est l'algorithme de DBSCAN. Pour cet algorithme, on définit le critère epsilon qui correspond à la distance entre deux points. Ensuite, l'algorithme définit le nombre de clusters. Dans notre cas, il a permis de déterminer 2 clusters.

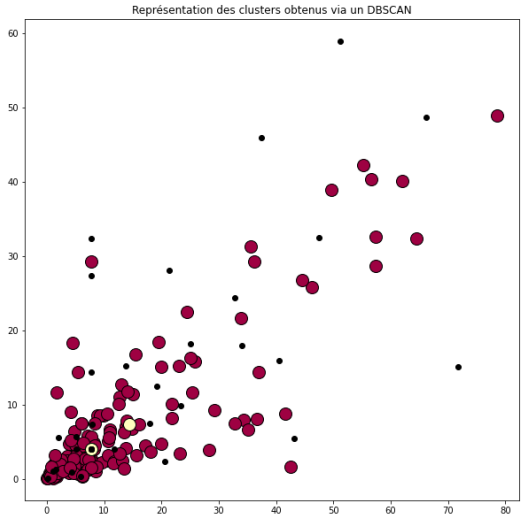


Figure 9. Graphique représentant les clusters du DBSCAN

La description des deux clusters sera détaillée dans l'analyse des modèles. Cependant, il semblerait que l'algorithme du DBSCAN soit le moins approprié à notre problématique pour discriminer des clusters de pays.

Analyses des modèles

Méthodologie de l'analyse

Pour faire une sélection de nos modèles et voir lequel discrimine le mieux les différents pays de notre échantillon, une analyse statistique va être effectuée. Pour cela, on va faire des comparaisons de moyennes entre nos différents clusters. Le schéma sera le suivant :

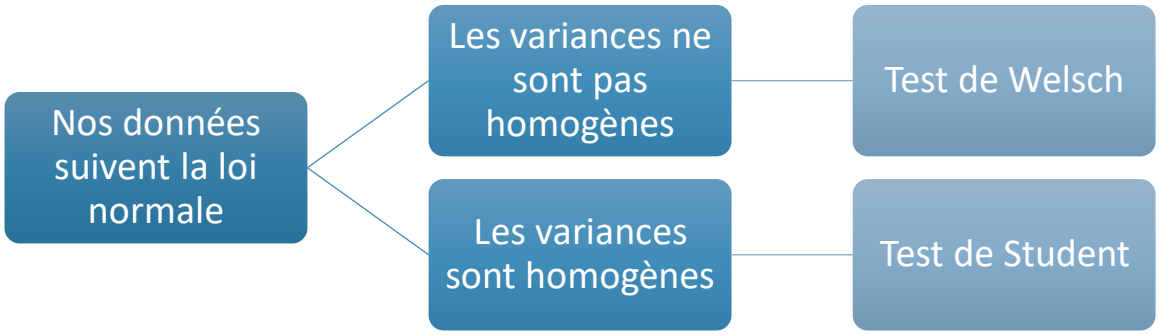


Figure 10. Schéma récapitulatif de la méthodologie de comparaison de moyennes

Analyse des algorithmes du K-Means

La première analyse va donc porter sur les résultats obtenus par l'algorithme du K-Means. On va donc voir si l'algorithme entraîné avec K = 2 permet d'obtenir des résultats pouvant séparer deux groupes de pays. Ensuite, on va s'intéresser à l'algorithme avec les paramètres : K = 4 ou K = 5. Le but ici sera de discriminer lequel sera le plus performant.

K-Means avec K = 2

Pour débuter notre analyse, on va essayer de faire une première séparation. La carte ci-dessous montre un découpage géographique de nos deux clusters.

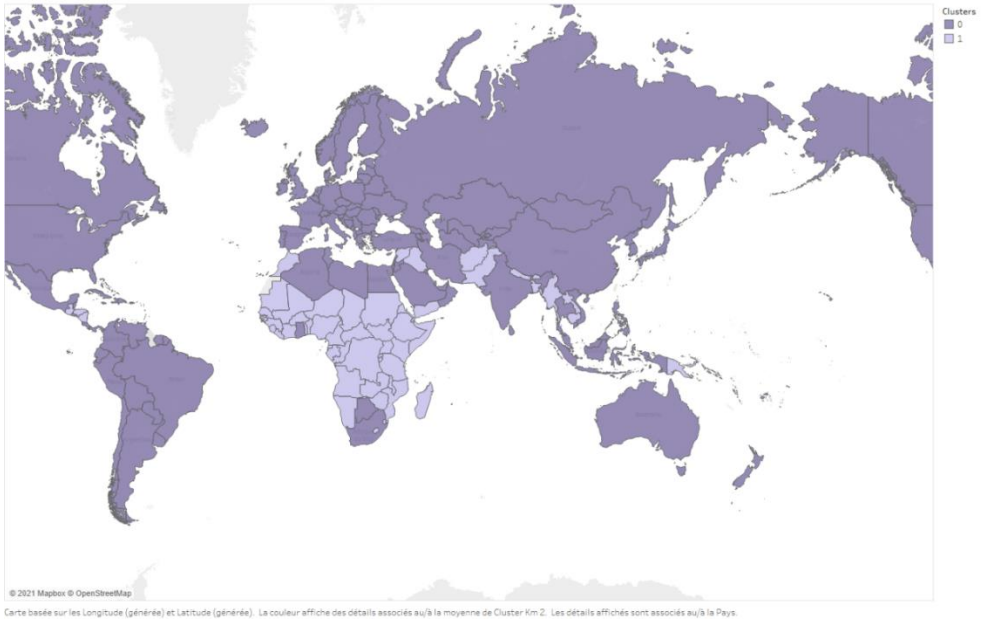


Figure 11. Carte représentant les pays appartenant aux clusters de l'algorithme K-Means avec K = 2

De nombreux pays d'Afrique font partie du cluster 1 ainsi que quelques pays d'Asie. Ensuite, nous allons nous pencher sur les moyennes à nos variables d'intérêt dans ces deux clusters.

Données éducatives

Pour plus de lisibilité, les données éducatives seront présentées sous la forme de deux graphiques. Les données en lien avec les ratios hommes/femmes ont été séparés des autres pour des questions de lecture.

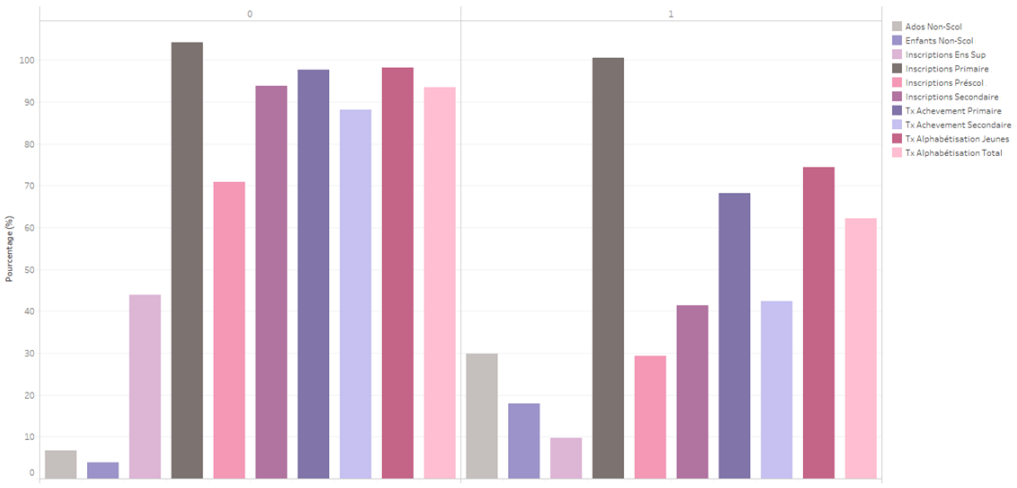


Figure 12. Scores moyens des différentes variables en lien avec l'éducation dans nos deux clusters

Les scores moyens obtenus à chacune de nos variables sur l'éducation sont significativement différents dans nos clusters. Le cluster 0 se caractérise par des faibles taux de non-scolarisation, de fort taux d'inscriptions quel que soit le cycle scolaire, des taux d'achèvement et des taux d'alphabétisation élevés.

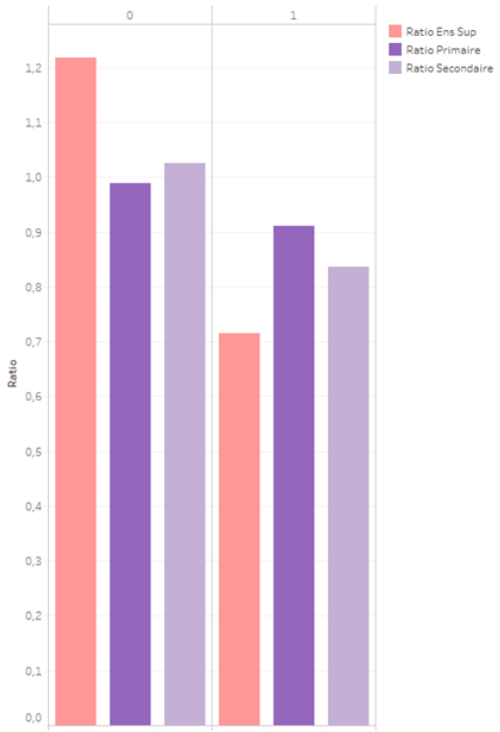


Figure 13. Ratios moyens de parité dans les différents cycles scolaires dans nos deux clusters

Concernant les ratios, on remarque que les pays du cluster 1 ont des ratio filles/garçons moins élevés que dans le cluster 0.

Données économiques

Les données économiques étant sur la même échelle, il n'y a qu'un seul graphique pour les représenter.

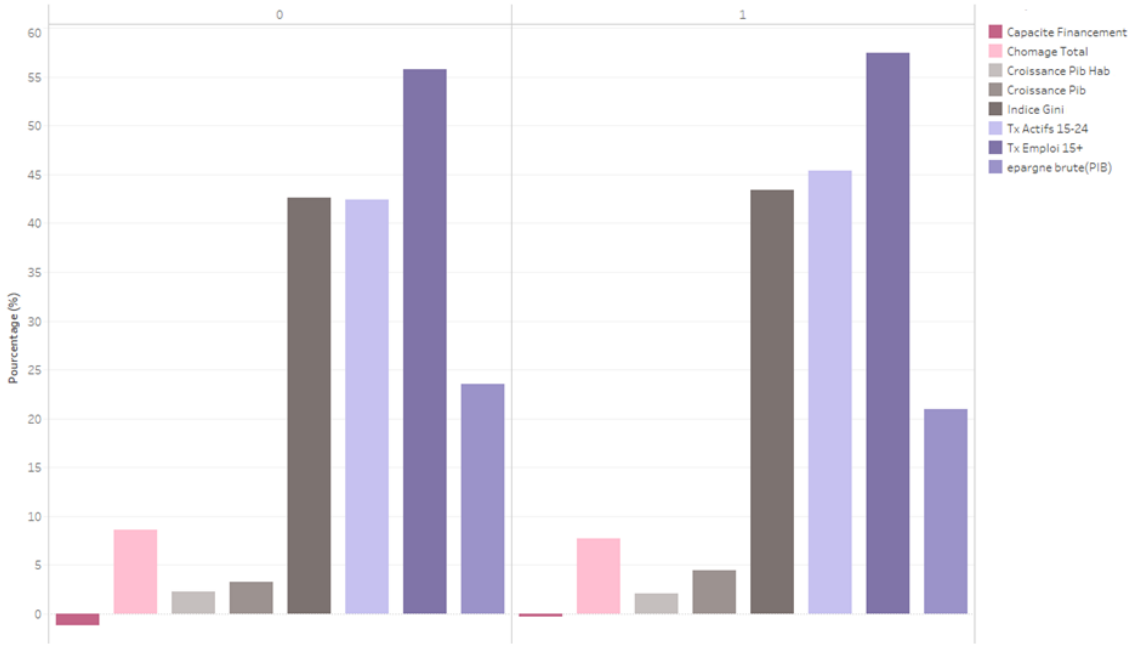


Figure 14. Scores moyens des variables économiques dans nos deux clusters

Les analyses statistiques montrent uniquement des différences statistiquement significatives pour les données en lien avec le chômage, le taux d'actifs et l'épargne brut. Les pays du cluster 1 ont un taux d'actifs entre 15 et 24 ans plus élevés que dans ceux du cluster 0.

Conclusion

Ce premier algorithme a perdu de déterminer deux clusters. Cette première segmentation nous a permis d'avoir deux groupes avec des différences statistiques significatives sur les données en lien avec l'éducation. En revanche, aucune donnée en lien avec la santé économique ne permet de différencier nos deux clusters sauf les données en rapport avec la population active et le chômage.

On remarque néanmoins qu'un taux de non-scolarisation élevé va de pair avec un taux d'emploi des 15-24 ans plus important. Cela semble être une première piste à explorer : les enfants non-scolarisés sont des enfants qui travaillent.

K-Means avec K = 4

Pour aller plus loin dans notre interprétation et notre compréhension des différences d'accès à l'éducation, un algorithme du K-Means a été entraîné avec cette fois-ci un K = 4. Les résultats de cette classification sont présentés dans la carte ci-dessous.

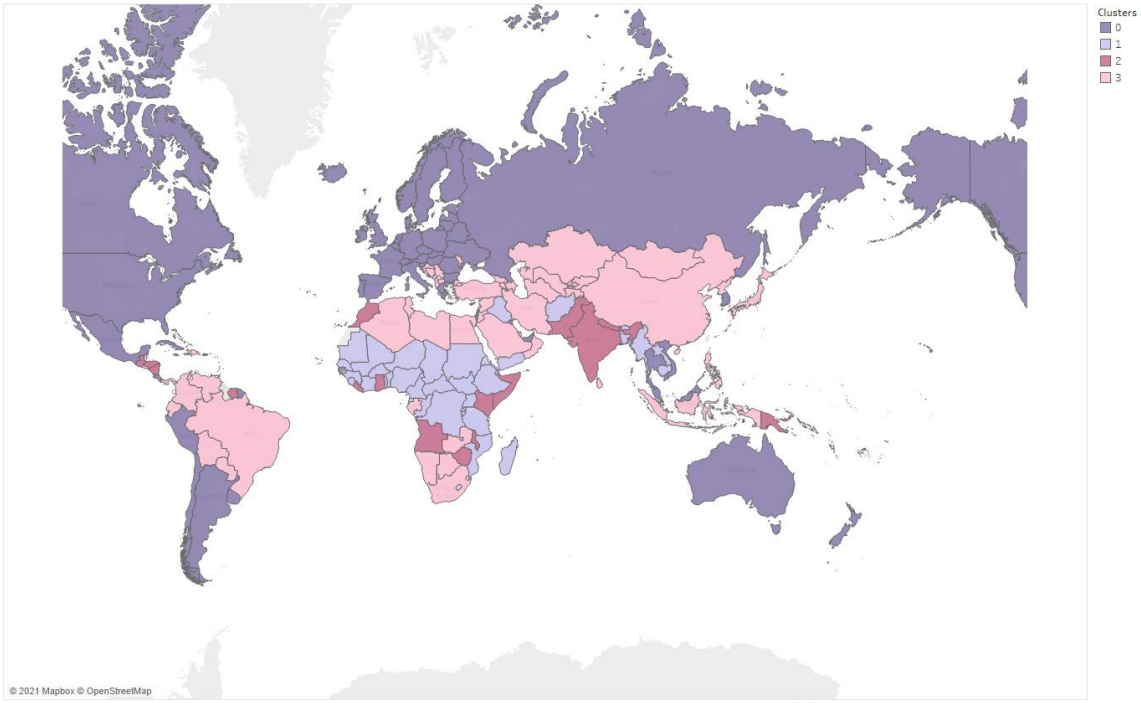


Figure 15. Carte représentant les pays appartenant à nos clusters de l'algorithme K-Means avec K = 4

La segmentation en quatre clusters permet d'avoir un niveau plus fin d'analyse de notre problématique. Le découpage nord/sud observé avec un K = 2 semble beaucoup plus complexe. On remarque que les pays d'Afrique, d'Amérique du Sud et de l'Asie se retrouvent distribuer parmi trois clusters. Pour mieux comprendre ce que cela signifie, on va s'intéresser aux scores moyens de nos variables.

Données éducatives

Là aussi les données éducatives sont séparées en deux graphiques pour faciliter la lecture des histogrammes.

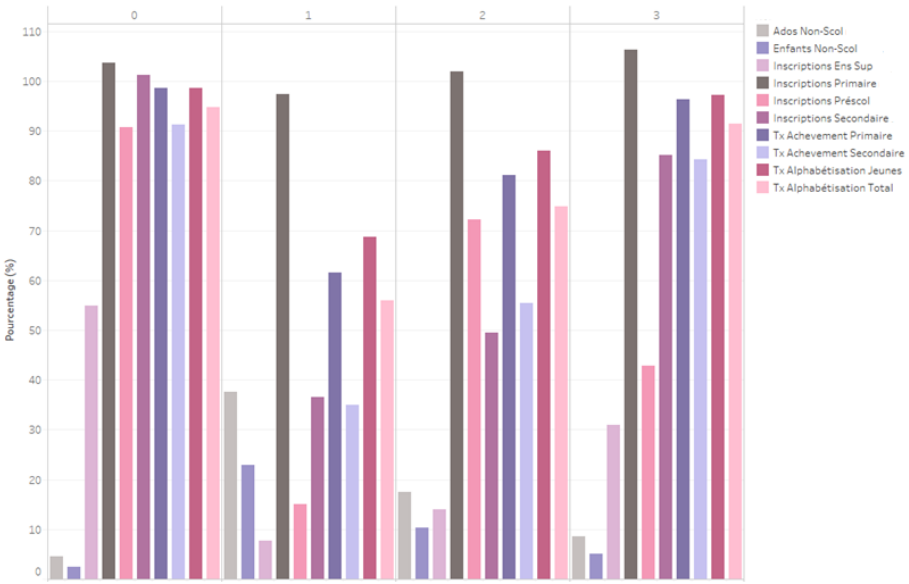


Figure 16. Scores moyens des différentes variables en lien avec l'éducation dans nos quatre clusters

Cette configuration de l'algorithme a permis de créer quatre clusters où les valeurs sur les données éducatives sont différentes. Les clusters 1 et 2 sont ceux qui présentent des valeurs importantes de non-scolarisation des enfants et des adolescents, ainsi que des taux d'inscription dans les cycles secondaires et d'enseignement supérieur plus bas. De plus, on observe que les taux d'alphabétisation sont plus faibles pour les pays de ces clusters.

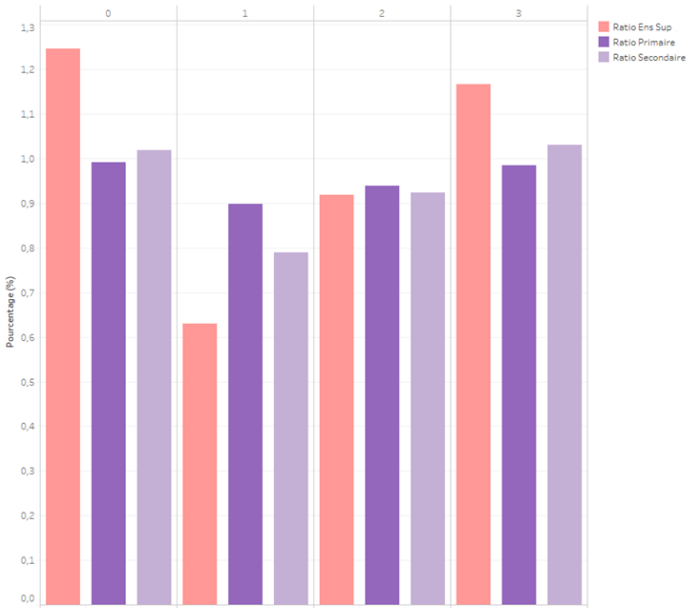


Figure 17. Ratios moyens de parité dans les différents cycles scolaires dans nos quatre clusters

Là aussi, les ratios dans les pays des clusters 1 et 2 sont statistiquement plus faibles que dans les pays des clusters 0 et 3.

Données économiques

Après avoir étudié les variables éducatives, on va se pencher sur les différences des variables économiques.

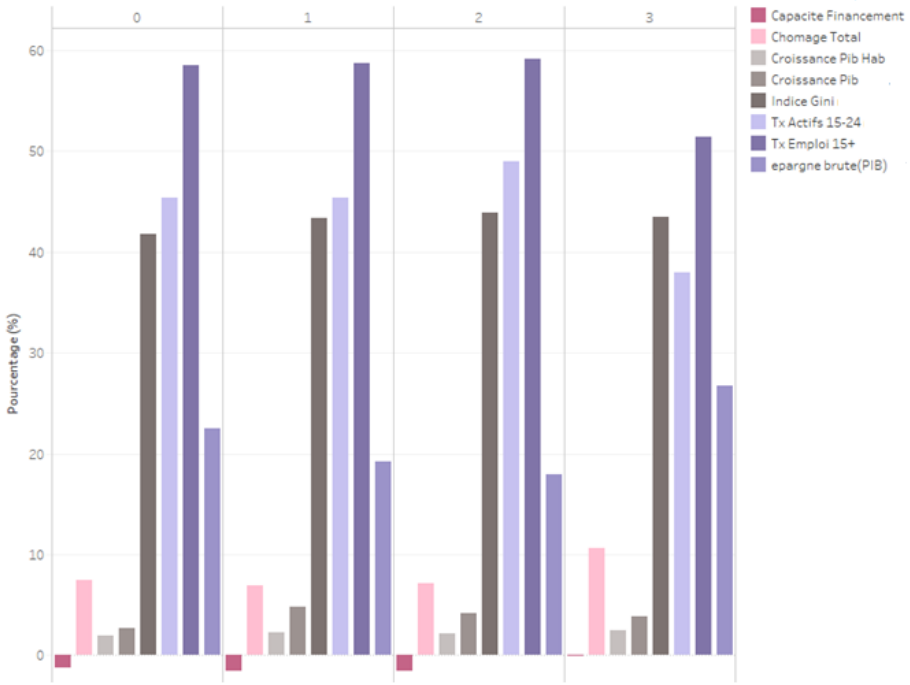


Figure 18. Scores moyens des variables économiques dans nos quatre clusters

Les clusters 1 et 2 se distinguent par leurs taux de chômage plus faibles, leurs croissances du PIB (total et par habitant) plus important et leurs taux de jeunes actifs plus élevés. Cependant, on remarque aussi que ces pays ont des besoins de financement plus importants.

Conclusion

Ce deuxième algorithme discrimine un peu mieux notre échantillon. On remarque que nos quatre clusters sont différents en ce qui concerne les données éducatives. Une partie de nos données économiques est différente entre chaque cluster, il s'agit essentiellement des données en lien avec le taux d'actif et la croissance du PIB annuel.

K-Means avec K = 5

La méthode du coude nous a montré que le nombre optimal pour K est de 4 ou 5. On va donc analyser les résultats obtenus avec K = 5.

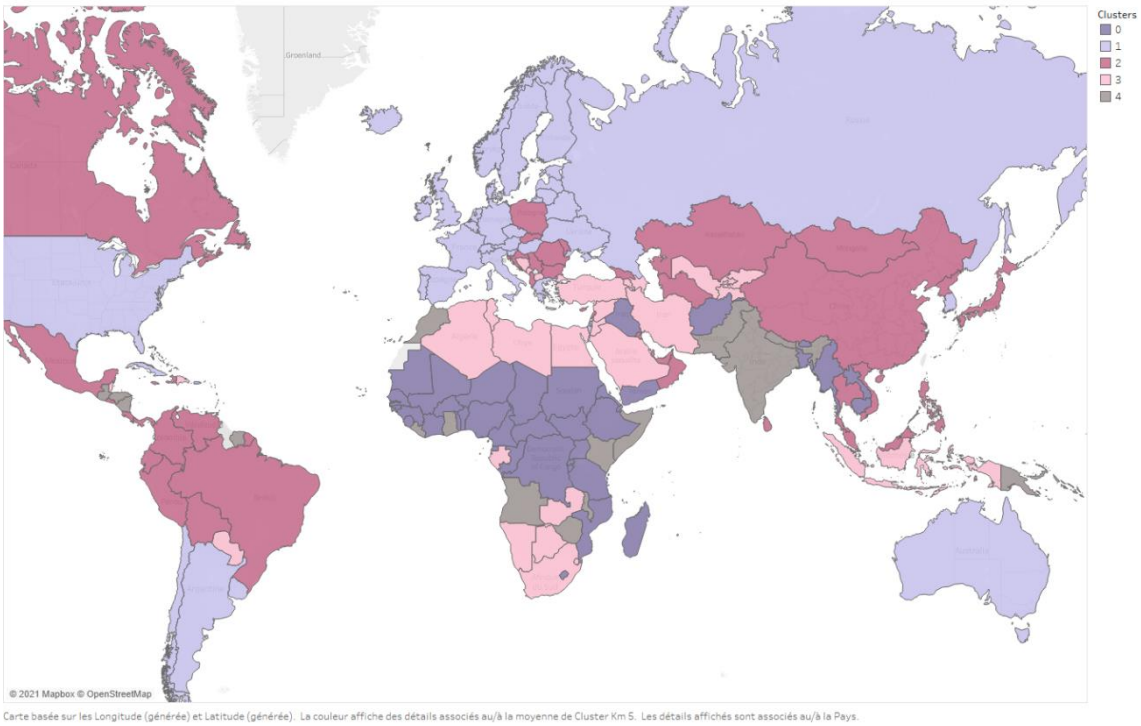


Figure 19. Carte représentant les pays appartenant à nos clusters de l'algorithme K-Means avec K = 5

Pour voir si cette segmentation est optimale, on va s'intéresser aux différentes variables.

Données éducatives

Là aussi, les données éducatives ont été divisés en deux parties.

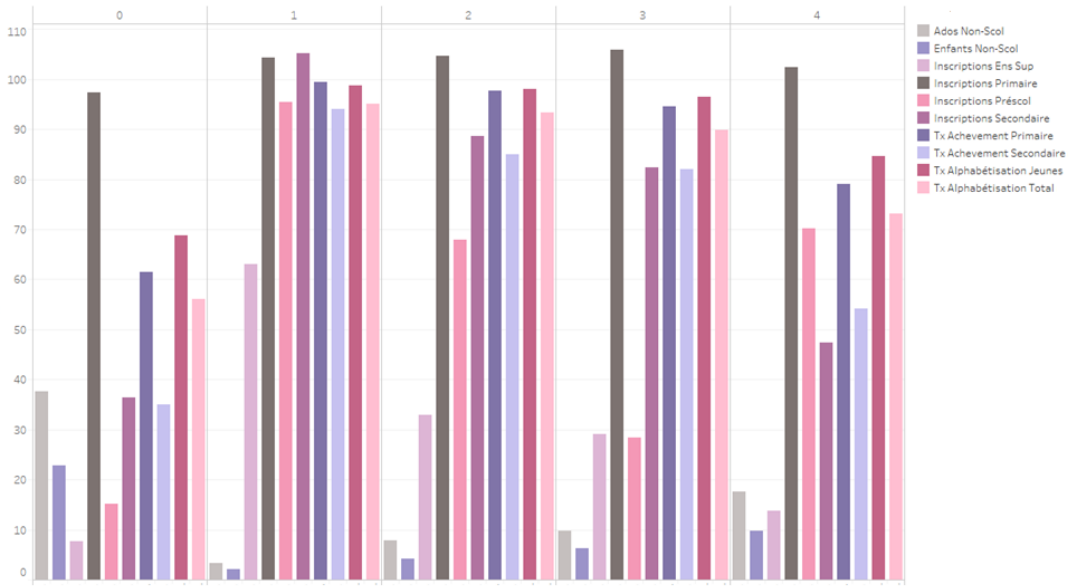


Figure 20. Scores moyens des différentes variables en lien avec l'éducation dans nos cinq clusters

Dans cette segmentation, les données éducatives discriminent moins bien nos clusters que dans l'algorithme du K-Means avec K = 4. La segmentation en 5 clusters a eu tendance à faire en sorte que nos scores moyens restent identiques.

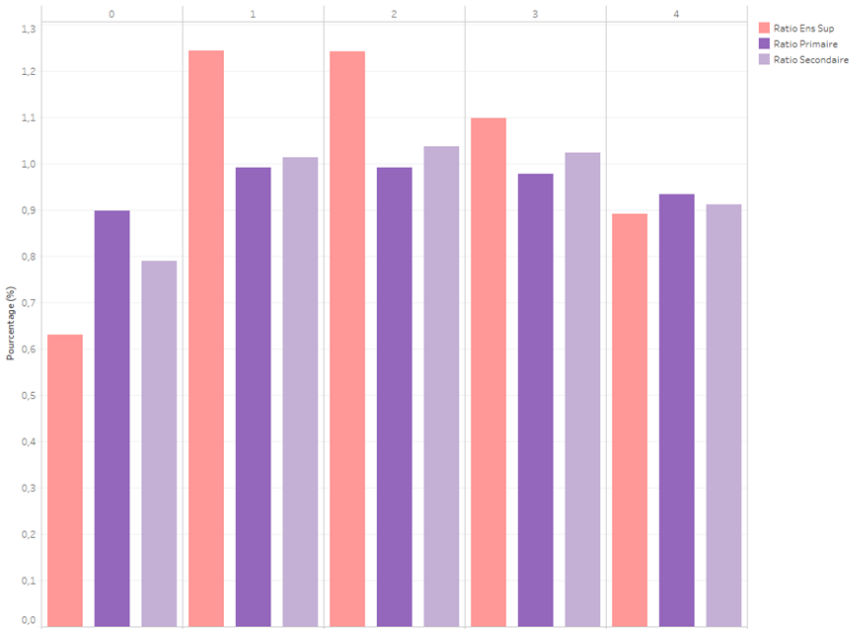


Figure 21. Ratios moyens de parité dans les différents cycles scolaires dans nos cinq clusters

Là aussi les ratios moyens de parité sont globalement identiques dans nos cinq clusters.

Données économiques

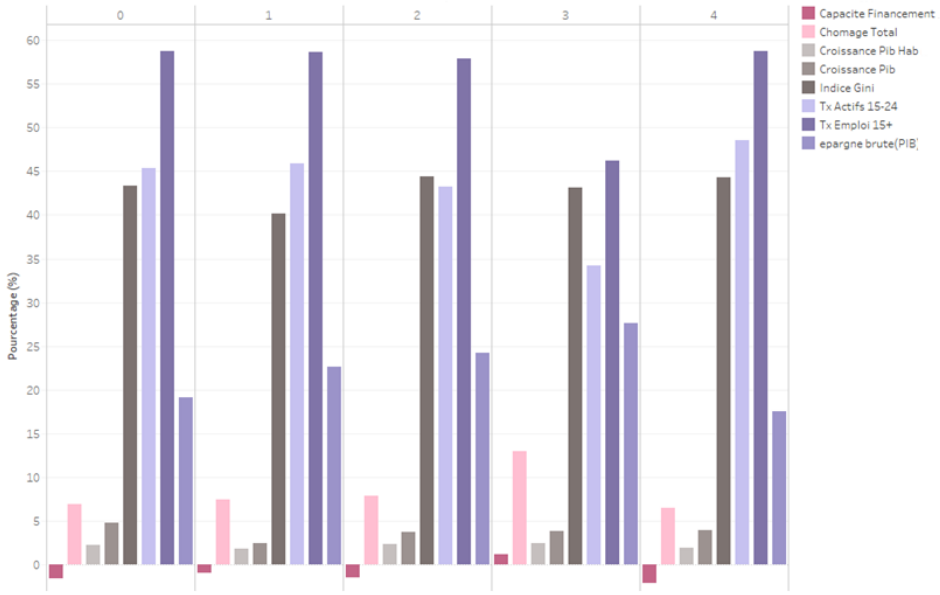


Figure 22. Scores moyens des variables économiques dans nos cinq clusters

Les données sur l'économie discriminent un peu mieux les cinq clusters mais les données sur l'éducation sont lissées.

Conclusion

Cet algorithme permet de discriminer les données éducatives mais il est moins performant que celui où K= 4. Cet algorithme n'est donc pas le plus performant pour notre problématique.

Analyse de la classification ascendante hiérarchique

Une autre méthode utilisée pour créer des groupes de pays : la classification ascendante hiérarchique. Pour cette classification, on a divisé notre échantillon en 3 clusters.

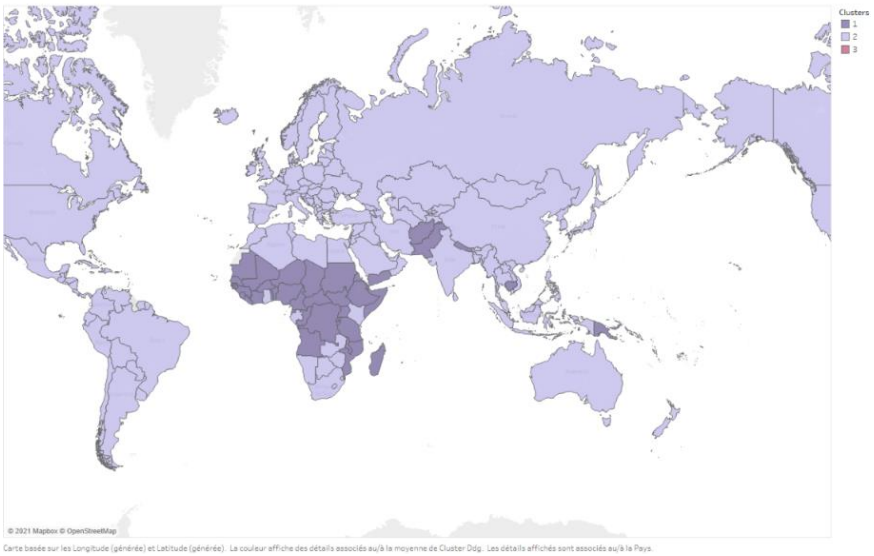


Figure 23. Carte des clusters obtenus avec la classification ascendante hiérarchique

Ce modèle nous permet bien d'obtenir trois clusters mais on observe que le cluster 3 contient un seul individu. Cela nous rapproche donc du découpage obtenu avec l'algorithme K-Means avec K = 2.

L'algorithme de DBSCAN

Cet algorithme sépare notre échantillon en 2 clusters. Le nombre de clusters n'est pas suffisant pour que notre modèle soit intéressé à utiliser.

Figure 24. Carte obtenue avec les clusters de l'algorithme DBSCAN

On ne va donc pas analyser les différents résultats obtenus.

Conclusion de l'analyse des modèles

Les différents algorithmes développés nous ont permis d'obtenir des segmentations intéressantes. Cependant, il semblerait que l'algorithme du K-Means soit celui qui fournissent des résultats intéressants. Dans tous les cas (K = 2 ; K = 4 et K = 5), nos clusters obtiennent des moyennes statistiquement différentes.

L'algorithme K = 2 était là pour surtout vérifier si on obtenait une réelle différence entre nos pays du nord et pays du sud. Dans l'algorithme où K = 5 on voit que les données en lien avec l'éducation sont moins discriminantes que dans l'algorithme où K = 4. Le modèle le plus performant dans notre étude semble être l'algorithme du K-Means avec K = 4.

Perspectives et pistes d'action

Synthèse de l'étude

Lors de cette étude, on a commencé par effectuer un état des lieux de l'éducation dans le monde. Les premières observations ont montré que les données de non-scolarisation, mais aussi d'accession et d'achèvement des différents cycles étaient globalement bons sauf dans certains pays d'Afrique où ces données montraient un faible niveau d'accès à l'éducation. Les données économiques ont montré des soucis de répartition des richesses dans certains pays.

Après avoir obtenu ces premières observations, on s'est intéressé à la création d'un modèle de segmentation qui nous a montré que l'algorithme le plus pertinent était l'algorithme du K-Means avec K = 4. En effet, cet algorithme nous permet d'obtenir quatre clusters bien distincts avec des différences significatives entre les différentes variables éducatives mais aussi entre les variables en lien avec l'économique, et plus particulièrement la croissance économique et la population active.

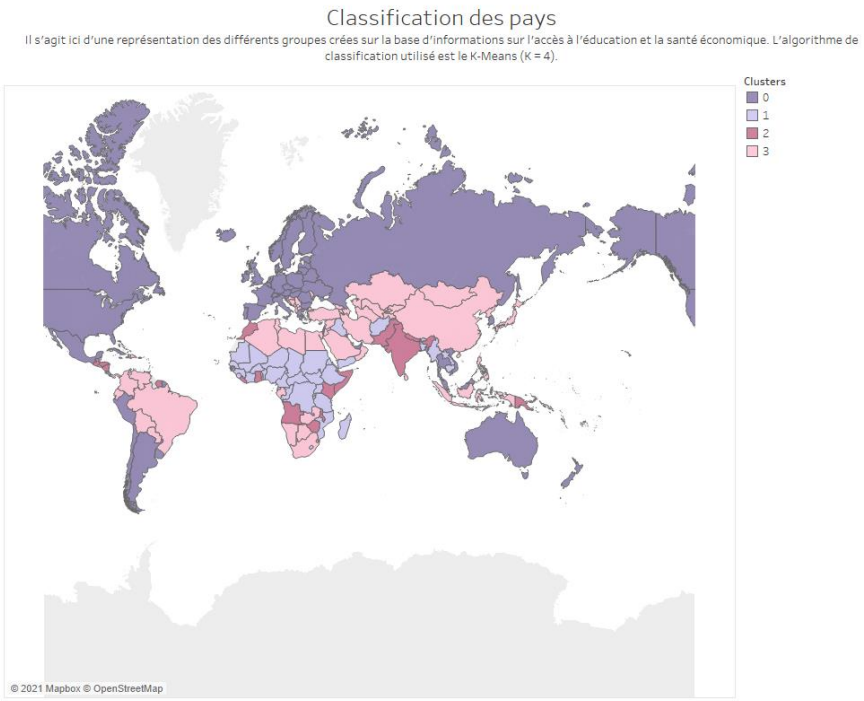


Figure 25. Dashboard de l'algorithme K-Means avec K = 4

Les pays du cluster 2 sont ceux qui ont les taux de non-scolarisation les plus élevés avec des besoins de financement importants. Néanmoins, ils ont aussi les taux de chômage les plus bas quelle que soit la tranche d'âge. La répartition des richesses reste équitable.

Les pays du cluster 1 ont des taux de non-scolarisation élevés mais qui tendent à se rapprocher de ceux des clusters 0 et 3. Cependant, leurs besoins de financement sont encore importants.

Caractéristiques des clusters à risque

Une fois notre segmentation réalisée, on va pouvoir se pencher un peu plus en détails sur les caractéristiques des pays des clusters 1 et 2.

Dans les pays du cluster 2

Les pays du cluster 2 sont ceux qui nécessitent une attention particulière. Parmi eux, on retrouve l'Inde, la Papouasie-Nouvelle-Guinée, le Maroc, la Somalie, le Kenya, le Zimbabwe ou encore le Suriname.

Ces pays sont d'anciennes colonies qui se retrouvent à devoir s'adapter à leur nouveau statut de pays indépendants. Ils se caractérisent par de fortes inégalités d'accès à l'éducation dû à un

système social différent (castes en Inde, enfants issus de milieu urbain vs milieu rural en Somalie ou Zimbabwe, etc.). La scolarité dans la plupart de ces pays est payante ce qui ne donne pas la même égalité d'accès à tous. Ces pays sont aussi marqués par de nombreux conflits (Papouasie-Nouvelle-Guinée, Somalie) où les enfants sont enrôlés très tôt dans les conflits. Enfin, il y a aussi de grandes inégalités d'accès en fonction du sexe. Les petites filles seront plus facilement vendues, mutilées ou mariées de force.

Dans les pays du cluster 1

Les pays du cluster 1 sont ceux où une action est aussi nécessaire mais elle sera différente de ceux du cluster 2. Le cluster 1 est composé essentiellement de pays d'Afrique et de la péninsule arabique comme le Tchad, le Yémen, la Syrie ou encore l'Afghanistan.

On remarque que ces pays sont des pays où des guerres civiles ont débuté dans les années 2000. Dans ces pays, la scolarité est obligatoire jusqu'à 14 ans en moyenne. Après cet âge, les enfants sont amenés à travailler dans des conditions très difficiles ce qui explique les taux de chômage peu élevés et les taux de non-scolarisation des adolescents importants. Les filles souffrent beaucoup plus de l'inégalité à l'éducation que les garçons. Elles sont souvent vendues ou mariées de force.

Mode d'action

Selon le cluster d'intérêt, le mode d'action sera donc différent.

Dans les pays du cluster 2, il semblerait que des mesures d'accompagnement soient le plus efficace. En effet, il s'agit de pays en transformation qui se retrouvent à devoir gérer leur indépendance acquise depuis une quarantaine d'années. Les années de colonisation ne les ont pas préparés à gérer leur économie de manière à savoir où placer leur fond. Ce sont des pays avec des ressources qui ont besoin de conseils et d'aides humaines pour pouvoir se développer.

Dans les pays du cluster 1, il y a avant tout un besoin d'intervention diplomatique. Les pays sont avant tout rongés par des guerres (civiles ou non) et ont besoin que les situations de conflits cessent. Une fois les conflits terminés, les droits des enfants pourront être rétablis et respectés.

Enfin, quel que soit le cluster, il semblerait qu'il est important de réduire les inégalités à l'accès à l'éducation. En effet, on retrouve aussi bien des inégalités dus aux genres mais aussi dus à

l'origine de l'enfant. Un enfant issu de milieu rural aura moins de choix d'établissement scolaire qu'un enfant de milieu urbain.

Conclusion

Les algorithmes de machine learning ont permis d'établir une segmentation des pays pour pouvoir réfléchir à des pistes d'action en fonction des problématiques de chaque pays. Il semblerait que la segmentation en 4 groupes soit suffisante pour pouvoir trouver des pistes. Notre algorithme a permis de déterminer deux groupes à risque et d'identifier les problèmes communs de ces pays.

Il peut donc être intéressant de développer ce modèle en y insérant d'autres caractéristiques autres que la santé économique. On pourrait donc continuer ce travail en y ajoutant des variables en lien avec la culture ou encore la situation de conflit. Cependant, on remarque que malgré la non-utilisation de ces données, les variables économiques nous ont permis de déterminer les problèmes de ces pays en y ajoutant quelques recherches sur le domaine.

L'analyse statistique et le machine learning ont permis de montrer que la problématique de l'accès à l'éducation dans le monde est quelque chose de complexe qui nécessite des solutions adaptées mais qu'il est possible d'utiliser les similarités de ces pays pour mettre en place des solutions.

Bibliographie

Banque Mondiale. (s. d.). *World DataBank | Explorer Créer Partager*. Consulté le 5 avril 2021, à l'adresse https://databank.banquemondiale.org/home.aspx

Deutsche Welle (www.dw.com). (s. d.). *Tchad : faible taux de scolarisation des filles*. DW.COM. Consulté le 1 avril 2021, à l'adresse https://www.dw.com/fr/tchad-faible-taux-de-scolarisation-des-filles/a-50788705

Enfants de Somalie. (2020, 1 octobre). Humanium. https://www.humanium.org/fr/somalie/

Enfants de Syrie. (2020, 10 octobre). Humanium. https://www.humanium.org/fr/syrie/

Enfants du Suriname. (2020, 28 mai). Humanium. https://www.humanium.org/fr/suriname/

Grâce aux « Écoles en boîte », les enfants de Papouasie-Nouvelle-Guinée peuvent enfin retrouver le chemin de l'école, après des années de conflit. (2004, 8 décembre). UNICEF. https://www.unicef.org/french/infobycountry/papuang_24481.html

Le système éducatif en Inde - Inde Éduc&Actions. (2018, 11 avril). Inde Éduc'Actions. https://inde-education-actions.org/civilisation/le-systeme-educatif-en-inde/

Pakistan : Des filles privées d'éducation. (2020, 28 octobre). Human Rights Watch. https://www.hrw.org/fr/news/2018/11/12/pakistan-des-filles-privees-deducation

Partenariat mondial pour l'éducation. (s. d.). *Zimbabwe | Partenariat mondial pour l'éducation*. Consulté le 1 avril 2021, à l'adresse https://www.globalpartnership.org/fr/where-we-work/zimbabwe

Sirelo.fr. (2021, 24 février). *Le système d'éducation Marocain - Coûts et procédures d'inscription*. https://sirelo.fr/demenager-au-maroc/leducation-au-maroc/

Yémen : deux millions d'enfants ne sont pas scolarisés. (2019, 25 septembre). LEFIGARO. https://www.lefigaro.fr/flash-actu/yemen-deux-millions-d-enfants-ne-sont-pas-scolarises-20190925

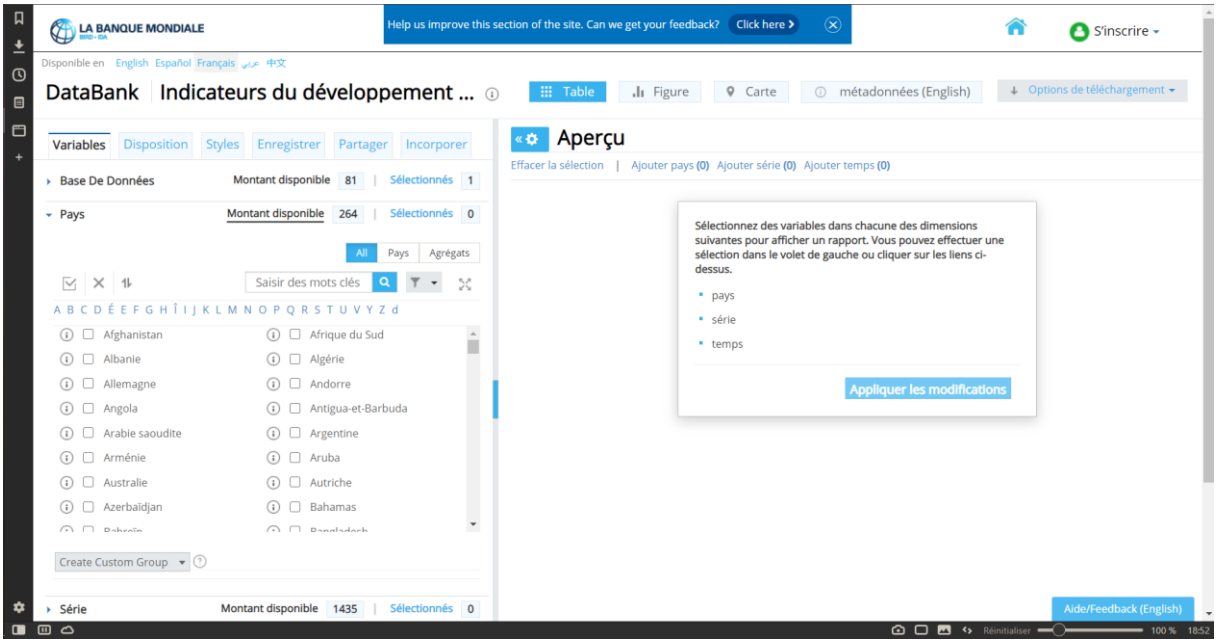
Annexe 1 : Notebooks

Les notebooks qui ont servi à réaliser ce projet sont au nombre de quatre :

- Un notebook contenant les étapes de nettoyage
- Un notebook contenant l'analyse descriptive exploratoire
- Un notebook contenant les algorithmes de Machine Learning
- Un notebook sur l'analyse des résultats de chaque modèle

Ces notebooks sont consultables sur GitHub : https://github.com/Sylvariane/acces_education

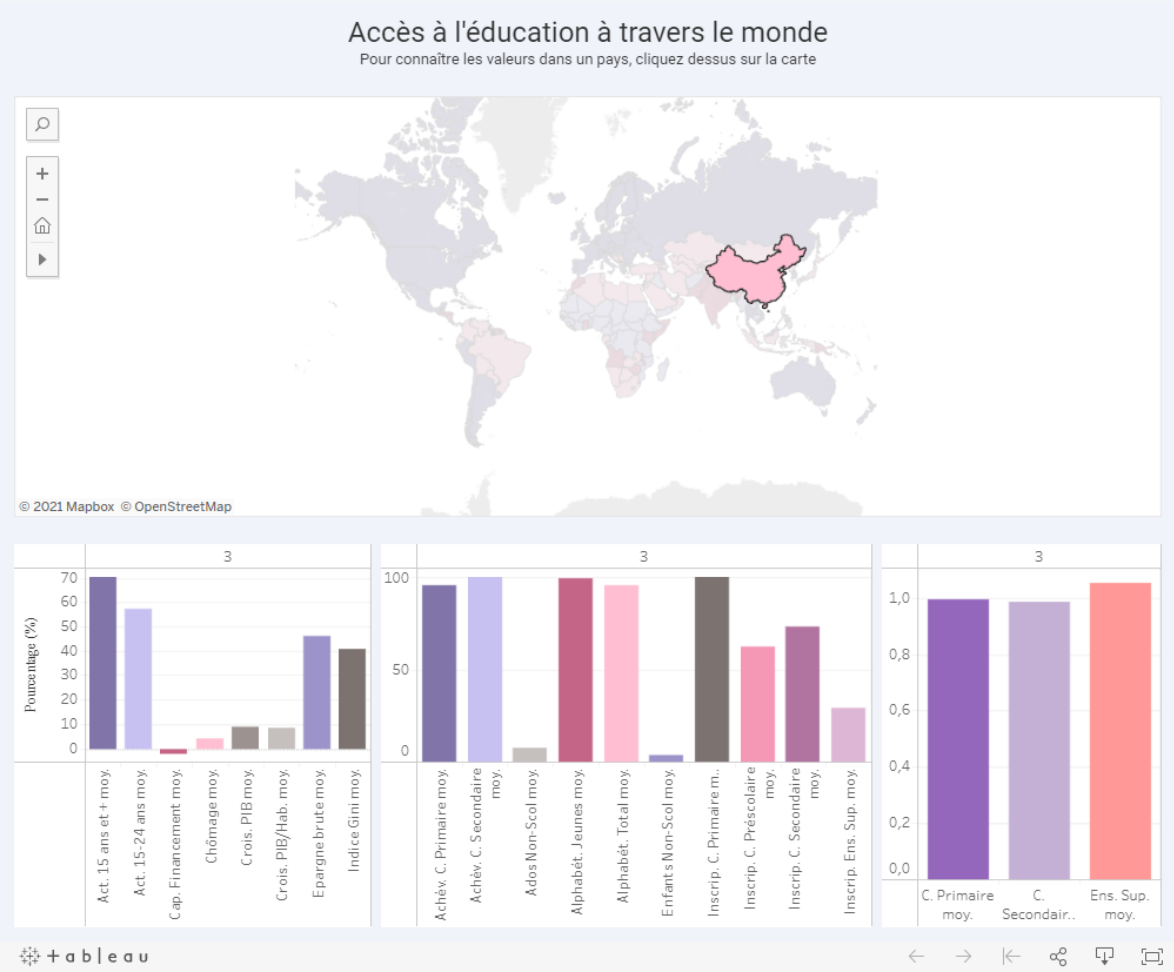
Annexe 2 : Outils de base de données



Les données utilisées proviennent d'une banque de données open source. Elles peuvent être donc complétées ou modifiées pour réaliser une étude sur une chronologie plus importante. Elles peuvent aussi être mise à jour pour entraîner notre modèle au fur et à mesure que le temps avance.

Annexe 3 : Dashboard

Un dashboard résumant toutes les données utilisées a été créée. Il est accessible à ce lien : https://public.tableau.com/views/Accesldducationtraverslemonde/Tableaudebord6?:language=fr&:display_count=y&:origin=viz_share_link



Les différentes feuilles qui ont servi à la construction de ce dashboard sont aussi disponibles à ce lien.