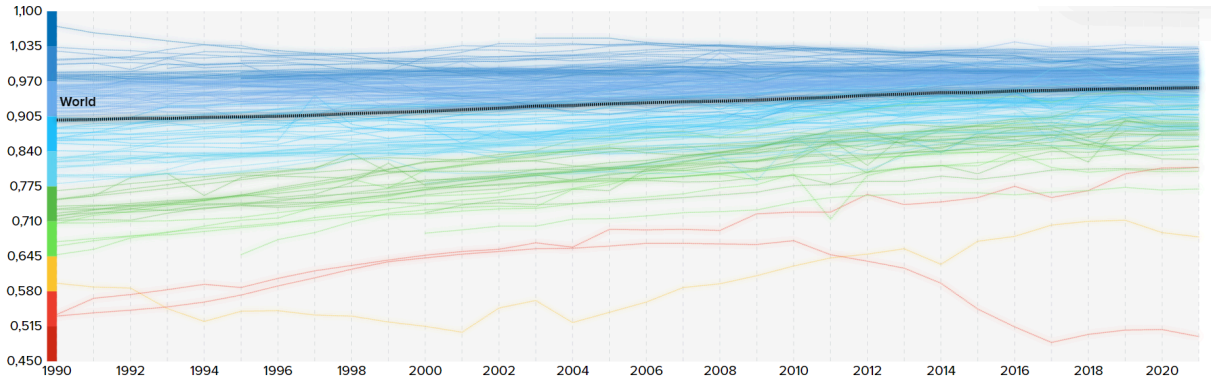
Lou Pelissolo
Tutrice: Dominique Benmouffek

Rapport - Projet d'Informatique S7

Ce projet d'informatique pour le semestre 7 avait pour objet la mise en place d'une plateforme de traitement de données statistiques, dans le but de mettre en évidence les biais de genre dans les données, inspiré par le livre *Femmes Invisibles* de Caroline Criado Perez. Pour ce faire, j'ai choisi d'utiliser des données sur deux indicateurs d'inégalités de genre calculés par l'ONU; le *Gender Development Index* (GDI) et le *Gender Inequality Index* (GII), et un indicateur de développement humain, l'IDH, dont on dispose de valeurs pour l'intégralité des pays du monde depuis 1990. Les trois catégories principales du calcul de ces indicateurs sont la santé, l'éducation, et le contrôle des ressources économiques, et c'est dans ces trois domaines que les critères de calcul diffèrent.

Le "Gender Development Index" (GDI)

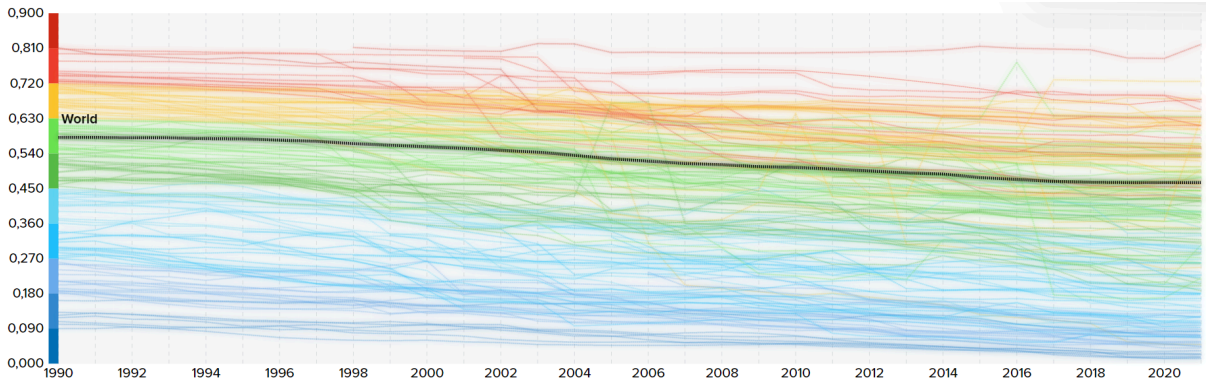


- Le GDI est une mesure qui compare le développement humain des femmes et des hommes dans un pays. Il est conçu pour montrer les écarts entre les genres dans les réalisations en termes de santé, de connaissances et de niveau de vie.

- méthode de mesure : Le GDI est calculé en prenant le rapport des indices de développement humain (IDH) pour les femmes par rapport à ceux des hommes. Les composantes de l'IDH sont l'espérance de vie, l'éducation (années de scolarisation moyennes et attendues) et le revenu national brut par habitant.

- Il s’interprète de la façon suivante: un GDI de 1 indique une parité parfaite entre les genres, tandis qu'une valeur inférieure à 1 indique des inégalités au détriment des femmes. La France possède par exemple un GDI de 0.985 (2021).

Le “Gender Inequality Index” (GII)



- Le GII mesure également les inégalités de genre dans ces trois dimensions importantes : la santé reproductive, l'autonomisation et le marché du travail.

- méthode de mesure: Le GII est calculé en fonction de plusieurs indicateurs, notamment le taux de mortalité maternelle, la proportion de sièges parlementaires occupés par des femmes, la proportion de femmes et d'hommes ayant au moins une éducation secondaire, et la participation à la main-d'œuvre.

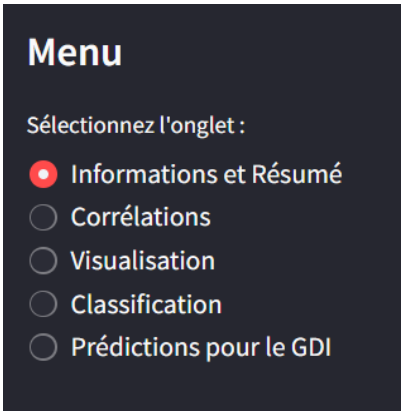
- Il s'interprète de la façon suivante; le GII varie entre 0 et 1, où 0 indique l'égalité des genres (aucune inégalité) et 1 une inégalité totale. La France possède un GII de 0.083, et est classée 22ème dans le monde.

J’ai retrouvé ces jeux de données sur le site du Programme des Nations Unies pour la Développement (UNDP), sous la forme de fichiers CSV. Le jeu de données GDI contient, pour divers pays, leur classement en termes d'IDH, le nom du pays, la valeur du GDI. Il fournit des mesures séparées pour les hommes et les femmes, telles que l'Indice de Développement Humain (IDH), l'espérance de vie, les années d'éducation prévues, les années

moyennes d'éducation, et le revenu national brut (RNB) par habitant, soulignant les écarts de genre dans ces indicateurs de développement.
Le dataset du GII inclut lui des informations par pays sur leur niveau de développement humain, la valeur du GII, le classement selon le GII, ainsi que des indicateurs spécifiques tels que la mortalité maternelle, le taux de natalité chez les adolescentes, le pourcentage de sièges occupés par des femmes au parlement, le taux d'éducation secondaire chez les femmes et les hommes, et la participation de la force de travail par genre. Ces données mettent en évidence les disparités entre les hommes et les femmes dans des domaines clés de la santé, de l'éducation et de l'empowerment politique.

J'ai également pu récupérer un dataset qui référence, par pays, toutes les valeurs de GDI de 1990 à 2021, ce qui nous permettra d'envisager de faire de l'apprentissage automatique sur ces valeurs.

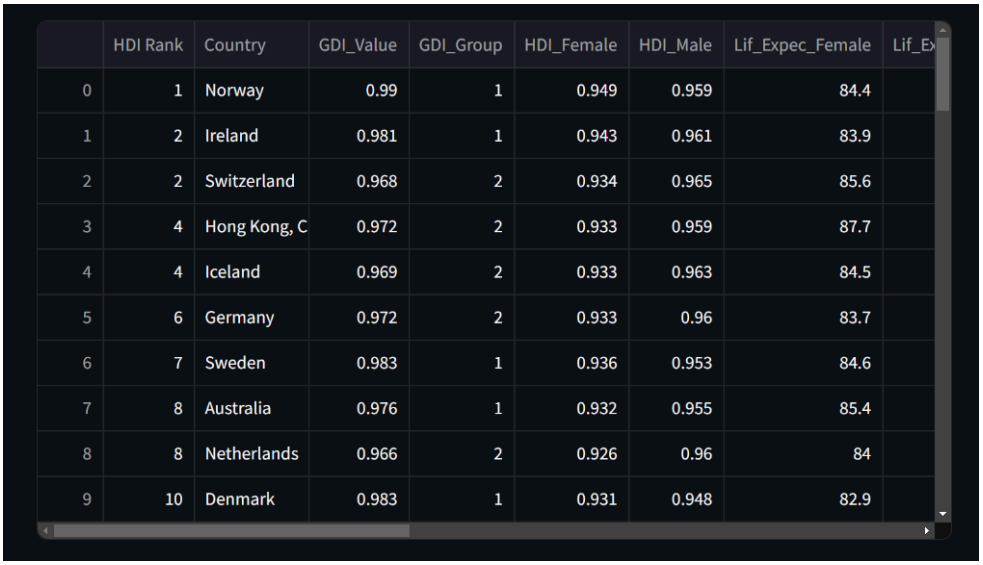
Pour créer l'interface de mon projet, j'ai choisi d'utiliser Streamlit, un outil open-source en Python spécialement conçu pour la création rapide d'applications web. Sa principale force réside dans sa simplicité et sa facilité d'utilisation, puisqu'elle utilise du code Python et fournit une large gamme de bibliothèques pour l'analyse de données, le machine learning et la visualisation, qui sont tous les outils dont j'avais besoin pour ce projet. De plus, et contrairement à R, Streamlit se distingue par sa capacité à créer rapidement des interfaces utilisateur élégantes.
Dans le cadre du projet, il me paraissait intéressant d'inclure cinq possibilités pour l'utilisateur, qui seront les cinq onglets du menu de la plateforme:



Ces cinq onglets sont les suivants: “Informations et Résumé”, “Corrélations”, “Visualisation”, “Classification”, et “Prédictions”. Ces différents onglets permettent chacun à l’utilisateur d’exploiter les données d’une manière différente, en fonction du fichier CSV qu’il a sélectionné.

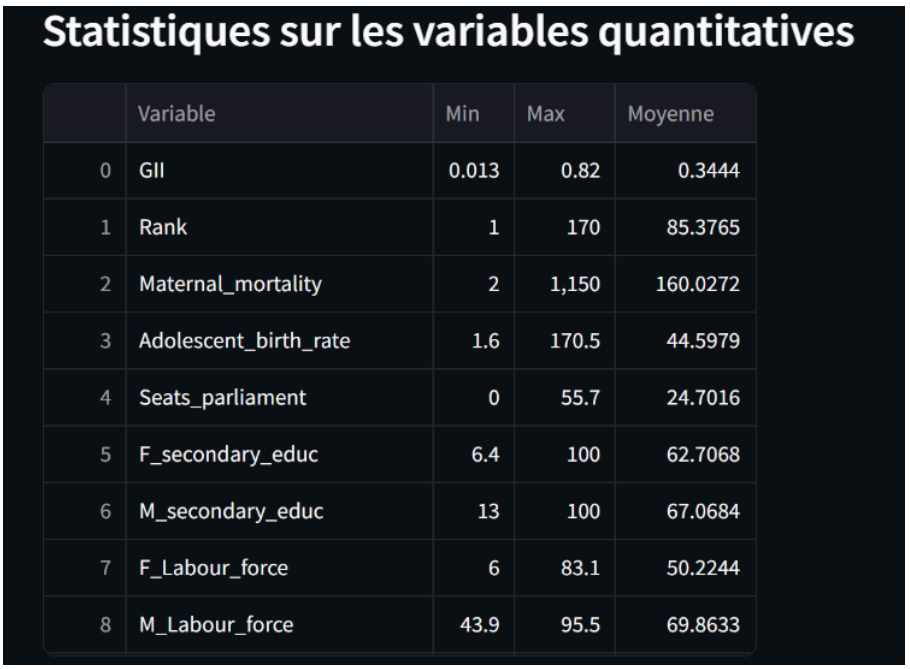
L’onglet Informations et Résumé

L’onglet “Informations et résumé” permet d’afficher les 70 premières lignes du jeu de données sélectionné. On peut ainsi avoir un aperçu rapide du fichier csv qu’on est en train de traiter, et des différents types de variables qui y sont exploités. Pour le fichier GDI par exemple, on a le tableau suivant, qu’on peut faire défiler;



	HDI Rank	Country	GDI_Value	GDI_Group	HDI_Female	HDI_Male	Lif_Expec_Female	Lif_Ex
0	1	Norway	0.99	1	0.949	0.959	84.4	
1	2	Ireland	0.981	1	0.943	0.961	83.9	
2	2	Switzerland	0.968	2	0.934	0.965	85.6	
3	4	Hong Kong, C	0.972	2	0.933	0.959	87.7	
4	4	Iceland	0.969	2	0.933	0.963	84.5	
5	6	Germany	0.972	2	0.933	0.96	83.7	
6	7	Sweden	0.983	1	0.936	0.953	84.6	
7	8	Australia	0.976	1	0.932	0.955	85.4	
8	8	Netherlands	0.966	2	0.926	0.96	84	
9	10	Denmark	0.983	1	0.931	0.948	82.9	

Ensuite, j’ai choisi d’intégrer des statistiques sur les variables quantitatives du tableau; j’ai calculé pour chacune d’elle son minimum, son maximum et sa moyenne. Pour le fichier GII, on a l’affichage du tableau ci-dessous.



	Variable	Min	Max	Moyenne
0	GII	0.013	0.82	0.3444
1	Rank	1	170	85.3765
2	Maternal_mortality	2	1,150	160.0272
3	Adolescent_birth_rate	1.6	170.5	44.5979
4	Seats_parliament	0	55.7	24.7016
5	F_secondary_educ	6.4	100	62.7068
6	M_secondary_educ	13	100	67.0684
7	F_Labour_force	6	83.1	50.2244
8	M_Labour_force	43.9	95.5	69.8633

Le tableau n'est pas forcément très explicite pour certaines des variables puisqu'il ne permet pas d'afficher les unités correspondantes, mais on peut par exemple apercevoir que, en 2021, le pays dans le monde ou les femmes étaient le moins scolarisées l'étaient à 6.4%, tandis que ce taux de scolarisation dans les pays moins inégalitaires pouvaient attendre 100%:

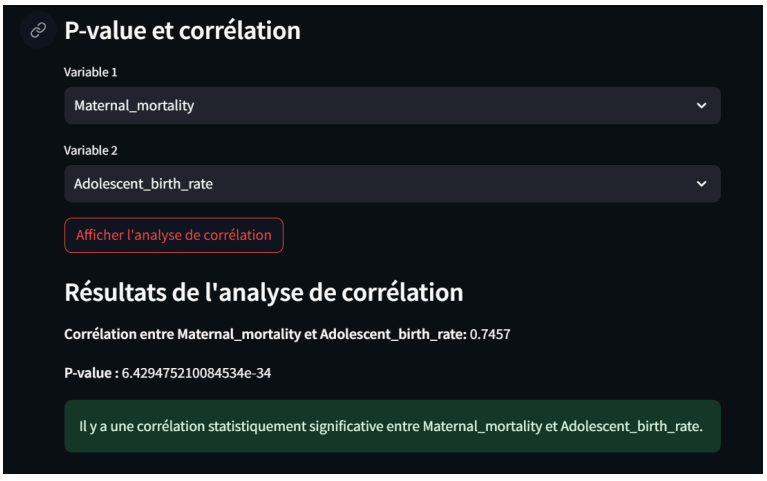
L'onglet Corrélations

L'onglet "Corrélations" est important car il permet de relier entre eux certains critères de calcul des inégalités. Pour cela, j'ai choisi de permettre à l'utilisateur de visualiser une matrice de corrélation entre toutes les variables quantitatives du jeu de données, ainsi que de réaliser une rapide analyse de corrélation.
La matrice de corrélation pour "Gender Inequality Index" nous donne l'affichage suivant:

Sur cette matrice, les cases coloriées en jaune et en violet indique une très forte corrélation entre les variables. Le jaune indique une corrélation positive; on peut dire par exemple que dans les pays ou le taux de mortalité maternelle est très elevé, le taux de natalité chez les adolescentes le sera aussi. A l'inverse, le violet indique une corrélation négative; par exemple, plus il y a de femmes scolarisées dans l'éducation secondaire dans un certains pays, plus le GII sera bas.

Des couleurs qui tournent vers le bleu, à l'inverse, indiquent qu'on ne peut pas relier l'évolution des deux variables entre elles.

Pour valider le niveau de corrélation j'ai choisi d'ajouter une fenêtre d'analyse, ou l'utilisateur peut sélectionner deux variables et obtenir un résultat sous la forme suivante;

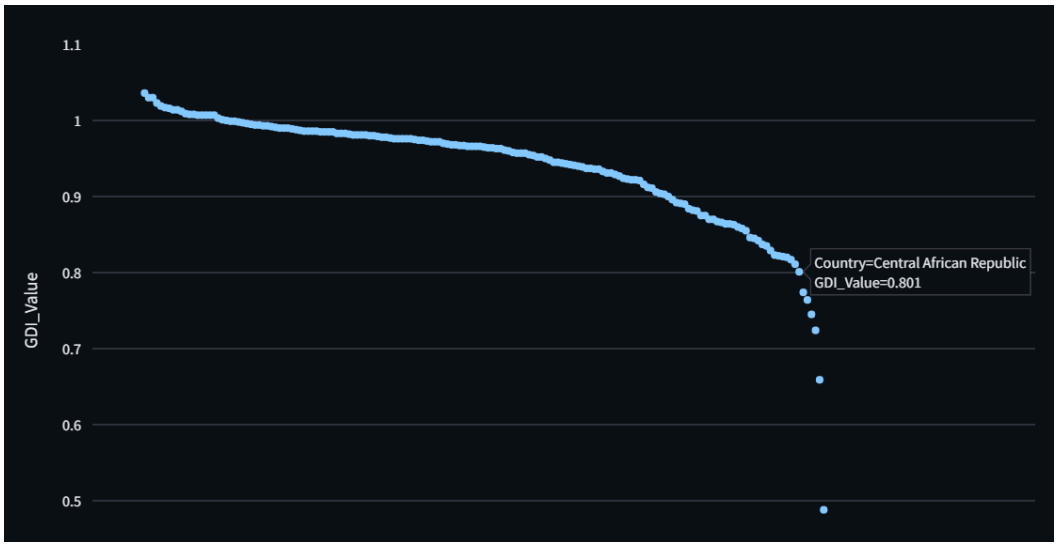


J’ai formaté la réponse de la plateforme pour qu’elle réponde positivement si la p-value est inférieure à un certain seuil fixé, et négativement si ce n’est pas le cas.

L’onglet Visualisation

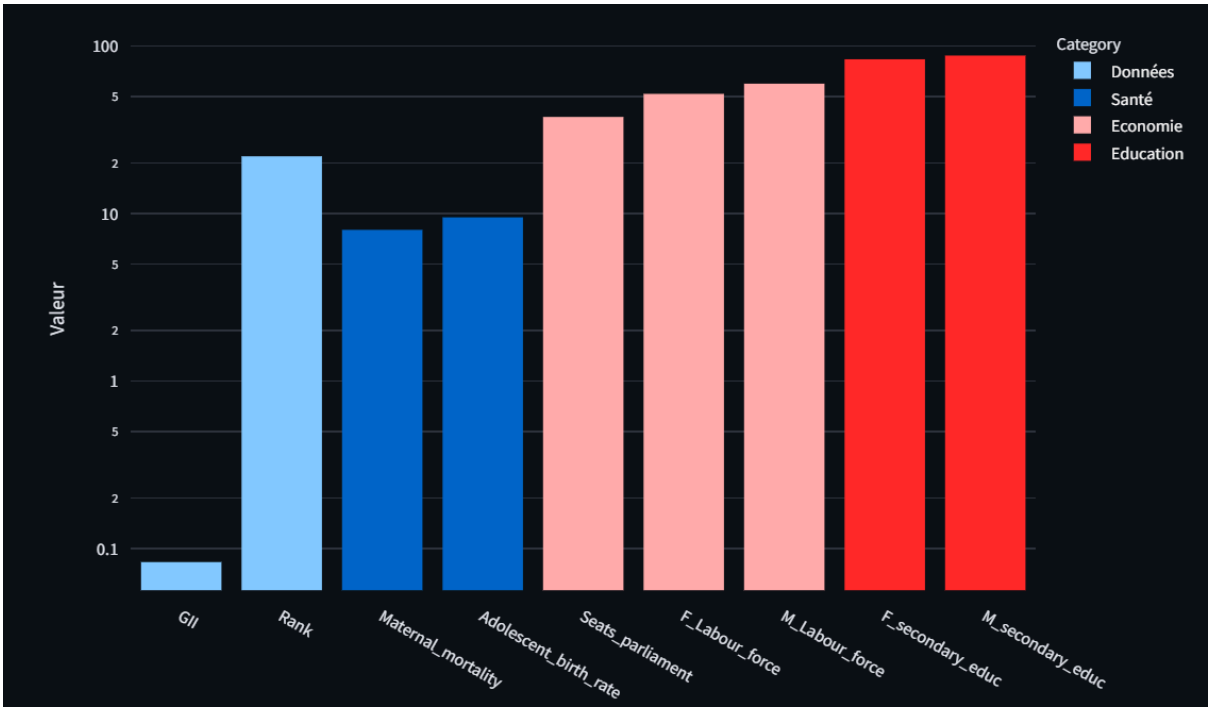
Cet onglet est essentiel puisqu’il permet à l’utilisateur d’avoir une représentation visuelle des données qu’il est en train de traiter. Il peut les voir de trois façons différentes: sous la forme d’un nuage de point interactif, sous la forme d’un histogramme par pays, et sous la forme d’une carte du monde.

Le nuage de point interactif est formé d’un point par pays du jeu de données, répartis sur le graphique en fonction de la valeur de leur indicateur (le GDI pour le graphique ci-dessous), et lorsque l’utilisateur passe la souris sur chaque point, il peut visualiser le label du pays correspondant.

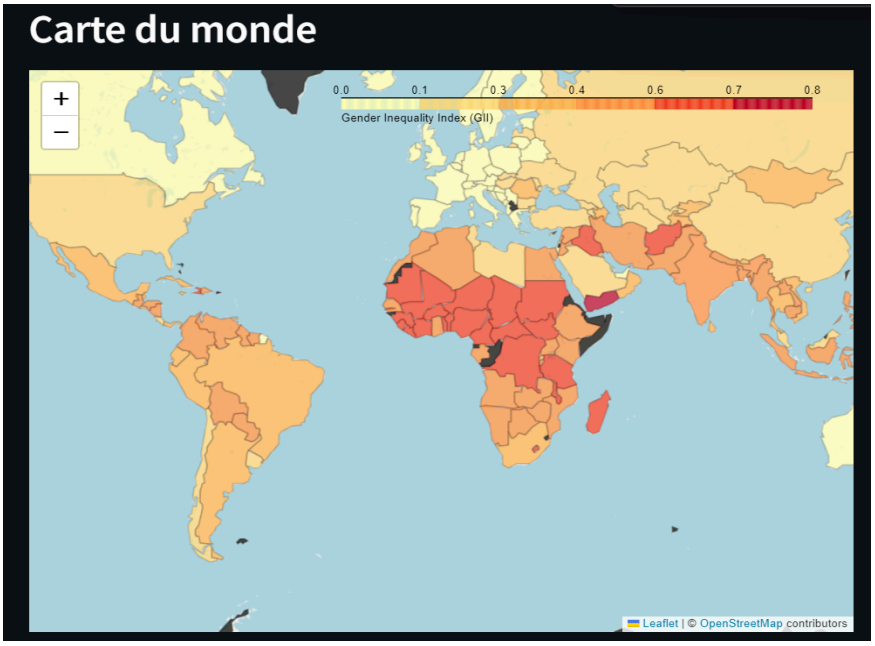


L’histogramme, lui, se concentre sur un pays en particulier choisi par l’utilisateur. Il lui permet de visualiser les performances du pays dans les trois catégories principales de calcul des indicateurs d’égalité entre les genres: la santé (en bleu marine), l’éducation (en rouge), et

le contrôle des ressources économiques (en rose). Puisque je souhaitais également afficher l'indicateur concerné sur l'histogramme, j'ai dû utiliser une échelle logarithmique. En effet, certaines variables sont exprimées en pourcentage, tandis que l'indicateur est une variable inférieure à 1; des soucis d'échelles se posaient donc. Dans le cas du GII de la France, on obtient l'histogramme suivant;



Finalement, l'aspect le plus efficace de la visualisation est de pouvoir se représenter l'ensemble des valeurs par pays sur une carte du monde coloriée. J'ai donc créé une échelle de couleur, allant du jaune pour les pays les plus égalitaires jusqu'au rouge pour les moins égalitaires, attribué une couleur à chaque pays selon son indice, et j'ai ensuite calqué le tout sur un fichier Géo JSON qui délimite les frontières des pays. On obtient finalement l'affichage suivant:



Les pays coloriés en noirs sont soit des pays pour lesquels l'UNDP n'a pas de données (le Groënland par exemple) soit des pays que je n'ai pas réussi à calquer sur le fichier Géo, à cause notamment de différences de dénominations entre les pays.

L'onglet Classification

L'ajout de cet onglet m'a paru intéressant pour reconnaître d'éventuelles similitudes sur les statistiques de genre entre les différents pays. J'ai donc choisi de classer tous mes pays selon 5 profils, en utilisant l'algorithme des K-Moyennes.

L'algorithme des K-moyennes est une technique de clustering qui organise un ensemble de données en k (ici k = 5) groupes distincts basé sur la similarité. Il commence par sélectionner aléatoirement k points comme centres de clusters initiaux. Ensuite, chaque donnée est attribuée au cluster le plus proche, selon la distance euclidienne. Après l'attribution, les centres des clusters sont recalculés en prenant la moyenne de toutes les données attribuées à chaque cluster. Ce processus est répété jusqu'à ce que les centres des clusters ne changent plus, indiquant que l'algorithme a trouvé une répartition stable des données en clusters.
Dans mon cas, je les ai ensuite trié de sorte à ce que les numéros de ceux-ci correspondent au niveau moyen d'égalité dans le cluster; ainsi, le cluster 1 représente les pays où la moyenne

de GDI est la plus faible (la plus élevée pour le GII), et le cluster 5 celui où la moyenne est la plus élevée. Il est important de préciser cependant que l'algorithme des K-Moyennes a pris en compte les critères de calcul des indicateurs (taux de mortalité maternelle, pourcentage de scolarisation chez les femmes…) tout autant que la valeur de l'indicateur en lui-même. Cela permet donc de reconnaître des profils de pays. Par exemple, le Qatar et le Maroc ont des valeurs de GII extrêmement proches, mais plus de 0.2 d'écart de GDI; cela témoigne de similitudes dans le domaine de la "Santé Reproductive" mais de forts écarts par exemple dans la participation des femmes à la force de travail, ou leur présence en politique. Ainsi, ils ne seront pas classés dans le même cluster.

J'ai choisi de permettre la visualisation des clusters sous différentes formes.
Tout d'abord, l'utilisateur peut choisir le cluster auquel il veut avoir accès, et ainsi consulter un tableau qui contient tous les pays classés par l'algorithme K-Means dans ce cluster.

Pays dans le Cluster 5:

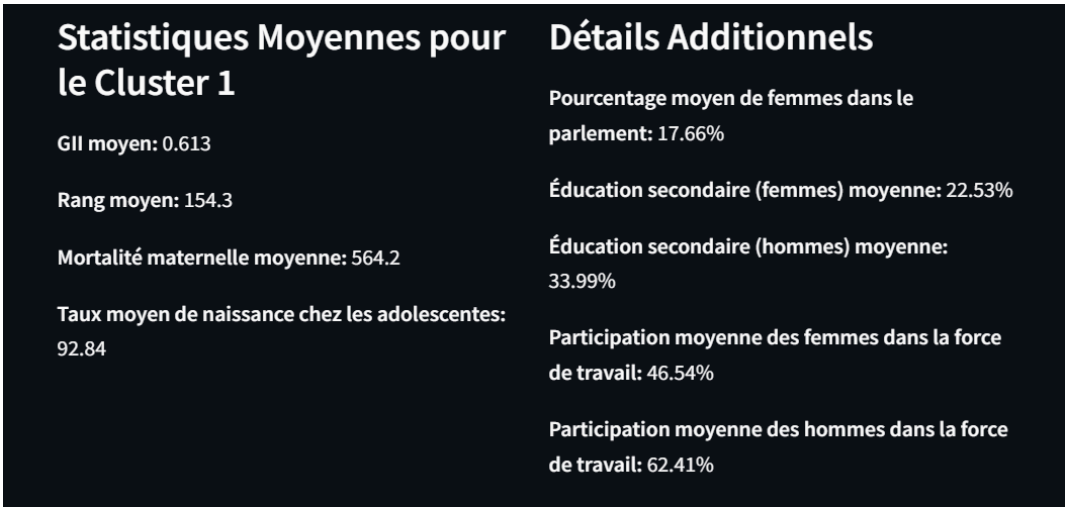
	Country	GII	Cluster
0	Switzerland	0.018	5
1	Norway	0.016	5
2	Iceland	0.043	5
4	Australia	0.073	5
5	Denmark	0.013	5
6	Sweden	0.023	5
7	Ireland	0.074	5
8	Germany	0.073	5
9	Netherlands	0.025	5
10	Finland	0.033	5

Pays dans le Cluster 1:

	Country	GII	Cluster
143	Eswatini	0.54	1
144	Equatorial Guinea	None	1
152	Congo	0.564	1
156	Papua New Guinea	0.725	1
157	Mauritania	0.632	1
158	Côte d'Ivoire	0.613	1
161	Togo	0.58	1
162	Haiti	0.635	1
163	Nigeria	0.68	1
167	Lesotho	0.557	1

Par exemple, le cluster 5 contient majoritairement des pays nordiques tels que la Suisse, la Norvège, la Finlande… Tandis que le cluster 1 est plutôt composé de pays d'Afrique subsaharienne tels que la République Démocartique du Congo ou le Togo.

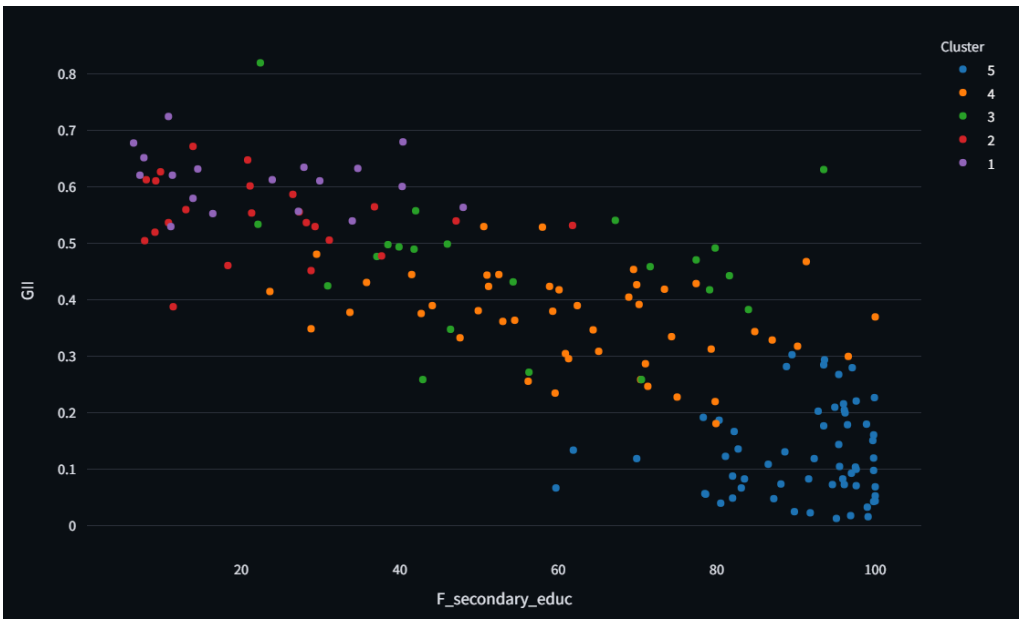
Pour le cluster sélectionné, l'utilisateur a également accès à des statistiques moyennes que j'ai calculé à partir des datasets pour chacun des clusters:

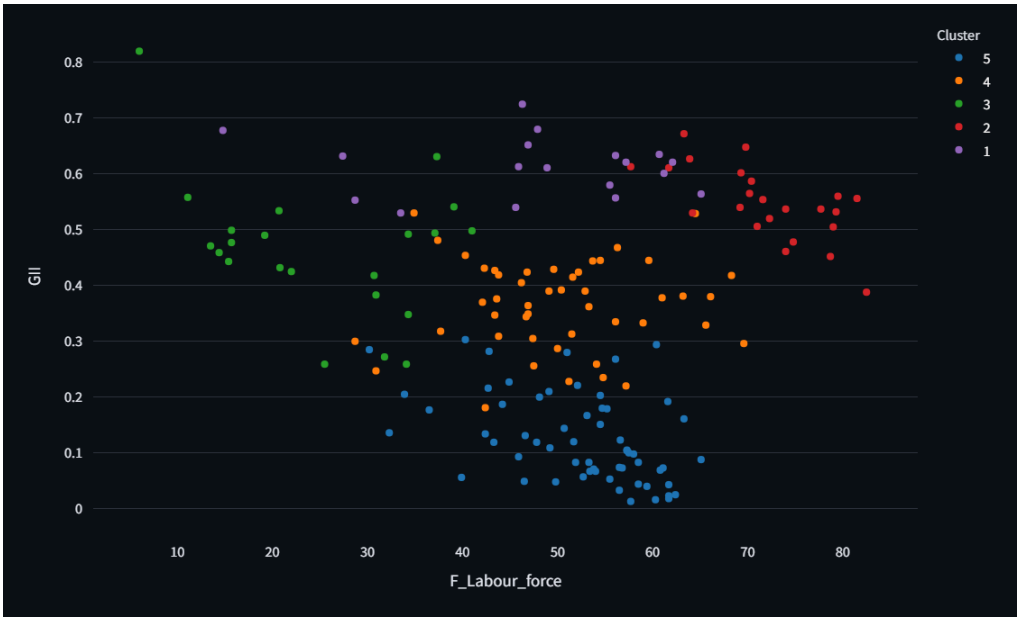


Finalement, on peut voir la répartition des clusters entre eux sous deux formes différentes: Tout d'abord, j'ai choisi de représenter un nuage de points des pays coloriés selon leur cluster, en laissant la possibilité à l'utilisateur de sélectionner la métrique de visualisation souhaitée parmi les colonnes du dataset.

On choisit par exemple de représenter la répartition des clusters selon le GII et le critère du taux de scolarisation des femmes dans l'éducation secondaire ("F_secondary_educ"). On voit que le partitionnement a été réalisé plutôt correctement selon ce critère. En effet, les pays du cluster 5 (en bleu) sont tous positionnés sur la droite en bas du graphique, là où le pourcentage de scolarisation est le plus élevé et le GII le plus faible, tandis que les pays des clusters 1 et 2 (en rouge et violet) sont situés en haut à gauche du nuage de point.

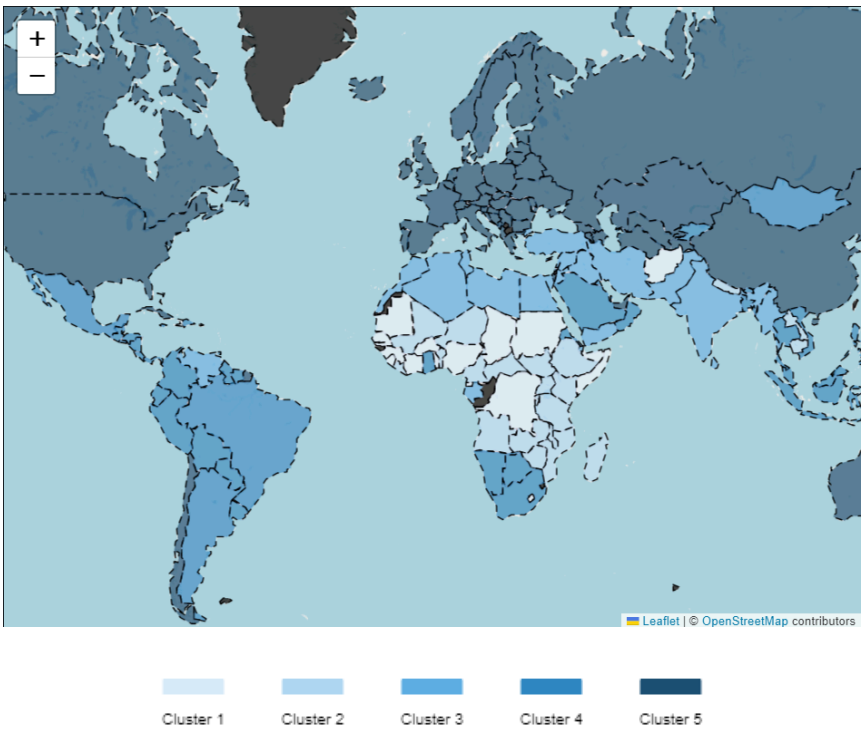
On remarque plus de dispersion pour les clusters 3 et 4, mais lorsque l'on choisit l'affichage selon d'autres critères, ils semblent être mieux différenciés (par exemple selon la participation des femmes à la force de travail "F_Labour_Force")





Finalement, j'ai également rajouté un affichage sur une carte du monde, de la même manière que dans l'onglet "Visualisation", mais cette-fois ci les pays sont coloriés en fonction de leur appartenance à un cluster.

Pour le GII, on obtient la carte suivante:



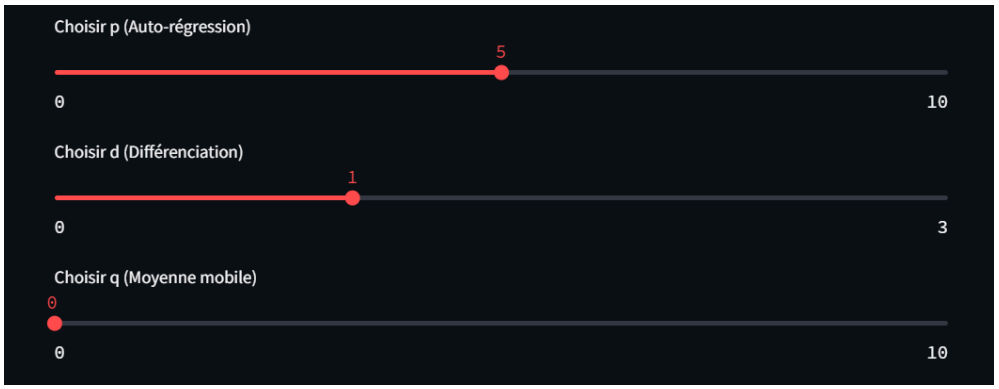
L’onglet Prédictions

Finalement, comme évoqué précédemment, j’ai pu recueillir des statistiques sur le GDI datées de 1990 jusqu’à aujourd’hui. Afin de pouvoir prédire d’éventuelles tendances dans les évolutions mondiales d’inégalités de genre, j’ai trouvé intéressant de réaliser de l’apprentissage automatique sur ces données, en entraînant mon modèle sur les 30 années précédentes. Après le test de plusieurs modèles d’apprentissage automatiques tels que Random Forest ou KNN, il m’est apparu que le modèle le plus pertinent dans le cadre de séries temporelles était le modèle ARIMA.

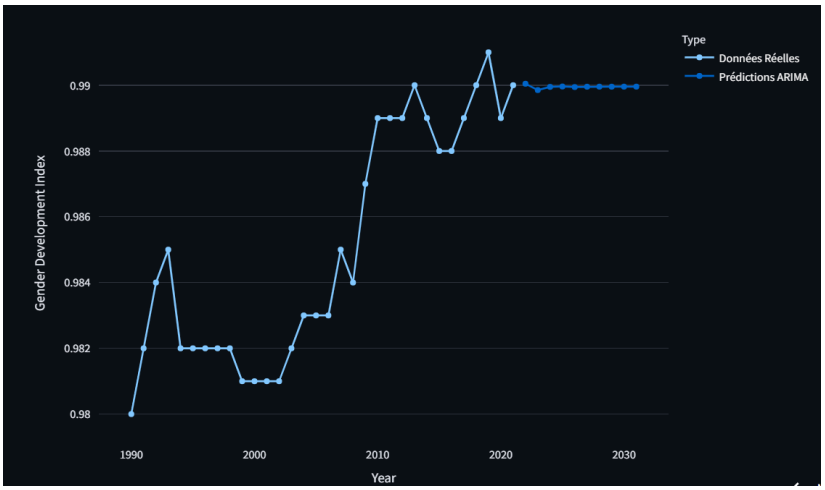
Le modèle ARIMA (AutoRegressive Integrated Moving Average) est une technique d'analyse et de prévision des séries temporelles qui combine trois composantes principales : autoregression (AR), différenciation (I pour Integrated) et moyenne mobile (MA). L'autoregression exploite la dépendance entre une observation et un nombre de ses retards. La différenciation aide à rendre la série temporelle stationnaire, c'est-à-dire à stabiliser la moyenne de la série sur le temps, en soustrayant une observation précédente d'une observation actuelle. La moyenne mobile modélise l'erreur de prédiction comme une combinaison linéaire des erreurs des périodes précédentes. Ensemble, ces composantes

permettent à ARIMA de capturer et de modéliser diverses structures dans les données temporelles, rendant ce modèle particulièrement puissant pour prévoir les tendances futures basées sur des observations passées. Le choix des paramètres pour ARIMA (ordres de l'autoregression “p” , de la différenciation “q”, et de la moyenne mobile “d”) est crucial et se fait souvent par analyse exploratoire.

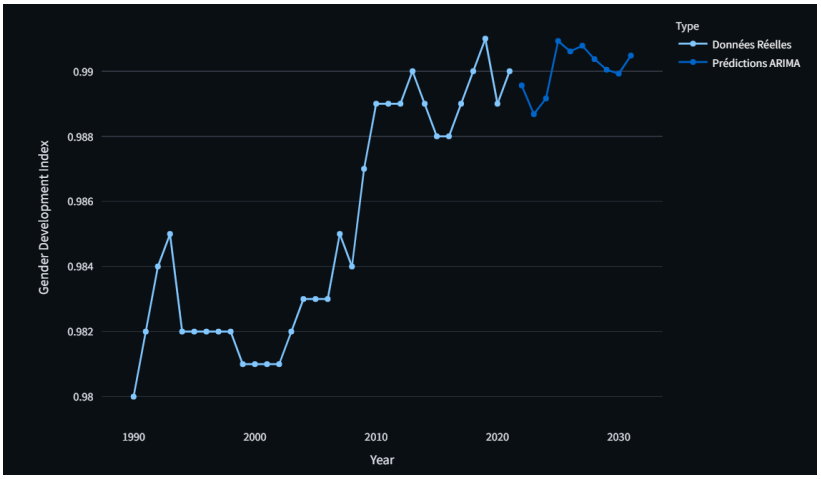
N’ayant pas eu le temps de définir pour tous les pays le triplet (p,q,d) de paramètres permettant d’obtenir la prévision la plus fiable, j’ai décidé de permettre à l’utilisateur de choisir lui-même ces indicateurs dans la plage de valeurs suivante:



Il peut ainsi voir l’influence de ces paramètres sur les prévisions. Pour la France par exemple, on voit que la variation du nombre de moyenne mobile n’a quasiment pas d’effet sur la courbe de prédiction, alors que l’augmentation des valeurs de p et d permettent d’obtenir plus de nuances dans les prédictions.



Prédictions ARIMA du GII Français avec (p,q,d) = (5,1,0)



Prédictions ARIMA du GII Français avec (p,q,d) = (10,3,0)

Cependant, comme on peut le voir pour de nombreux pays comme l'Afghanistan par exemple, la situation politique et économique du pays a un impact trop grand sur la situation des femmes pour pouvoir obtenir des prédictions fiables simplement avec de l'apprentissage automatique, qui ne permet pas de prédire par exemple une guerre ou l'arrivée d'un régime autoritaire au pouvoir.

Analyse des résultats

A présent, j'aimerai utiliser les affichages de ma plateforme Streamlit afin de mettre en lumière divers biais de genre dans les données sur la situation des femmes dans le monde. Les indicateurs d'inégalités de genre du GDI et du GII calculés par l'ONU sont les plus fiables et les plus utilisés de ceux dont on dispose aujourd'hui. Cependant, ils sont soumis à plusieurs critiques.

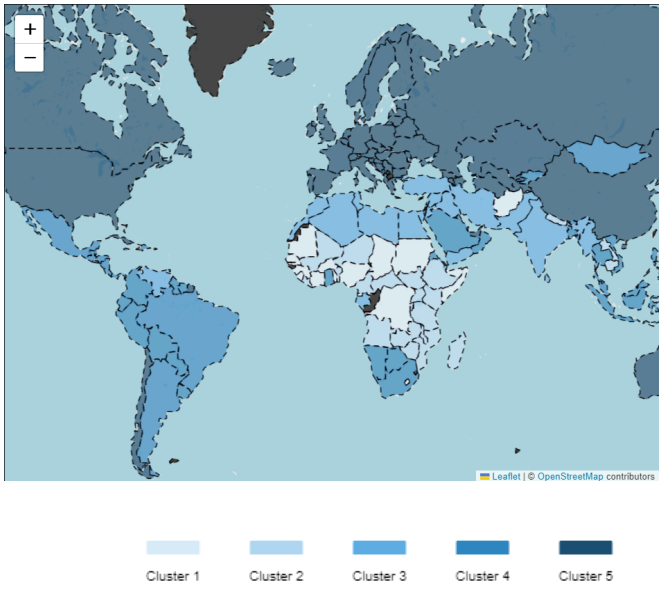
Tout d'abord, le calcul du GDI est déséquilibré puisqu'il se base sur deux critères problématiques:

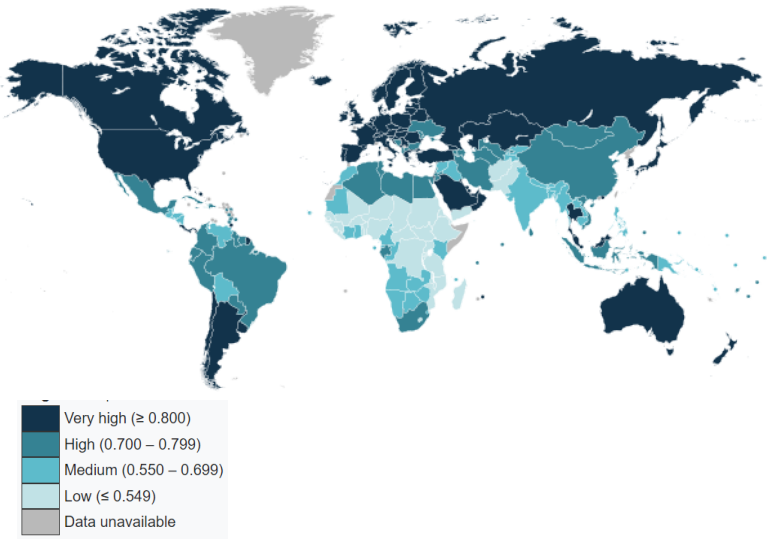
- Le GDI prend en compte l'écart d'espérance de vie entre les hommes et les femmes dans tous les pays du monde comme un facteur d'inégalité. Or, biologiquement les femmes ont une espérance de vie plus élevée que les hommes, ce qui ne constitue pas un facteur de non-discrimination.

- Le calcul se base également sur le Revenu National Brut par pays chez les hommes et chez les femmes, or cela fausse les données entre les pays pauvres et les pays riches, puisque les pays riches, avantagés par le haut niveau de revenu global, vont automatiquement apparaître comme plus égalitaires.

Le GII semble donc globalement être un meilleur indicateur, notamment grâce au fait qu'il prend largement en compte des facteurs de santé reproductive, même si le critère "Participation des femmes à la main d'oeuvre" est également faussé puisqu'il ne tient pas rigueur du plus faible salaire des femmes à travail égal, et du taux plus important de travail domestique chez les femmes.

Procédons maintenant à l'analyse des clusters de pays formés dans l'onglet Classification.

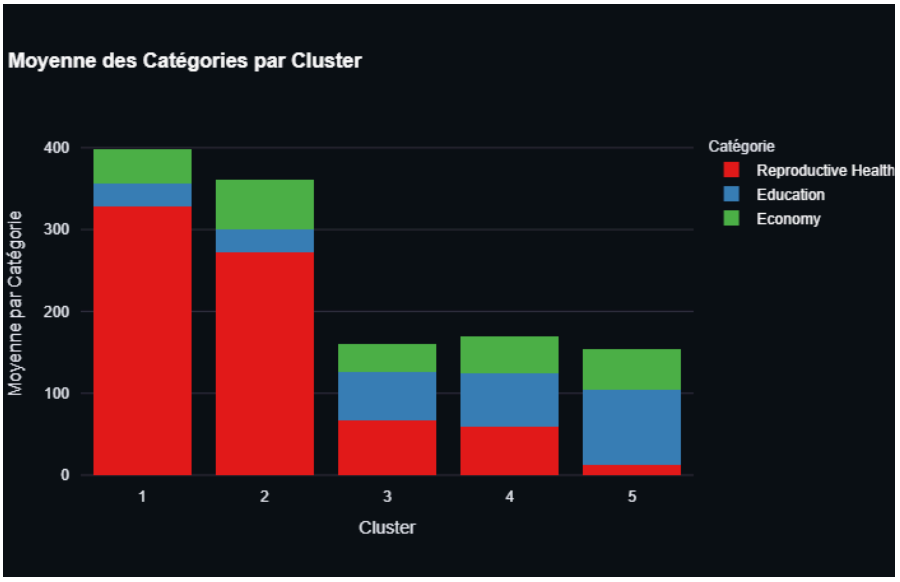




Carte des clusters obtenus avec l'algorithme K-means *Classification des pays selon leur HDI par l'ONU*

La carte mondiale colorée illustre le regroupement des cinq clusters de pays, créés par l'algorithme K-means, selon les critères du GII. En parallèle, la carte de l'UNDP (United Nations Development Programs) classe les pays selon leur Indice de Développement Humain, avec des catégories allant de très élevé (>0.8) à faible (<0.549).
On remarque qu'il y a une certaine similitude dans la distribution géographique des deux cartes ; par exemple, le cluster 5 semble englober tous les pays de l'OCDE, ayant donc un développement très élevé, tandis que le cluster 1 semble rassembler les pays que l'ONU

considère comme les moins développés. Cela mène à deux observations : premièrement, le GII apparaît comme un indicateur efficace pour regrouper les pays aux profils similaires via l'algorithme des K-moyennes, et deuxièmement, l'égalité des genres est très fortement corrélée au développement économique des pays.

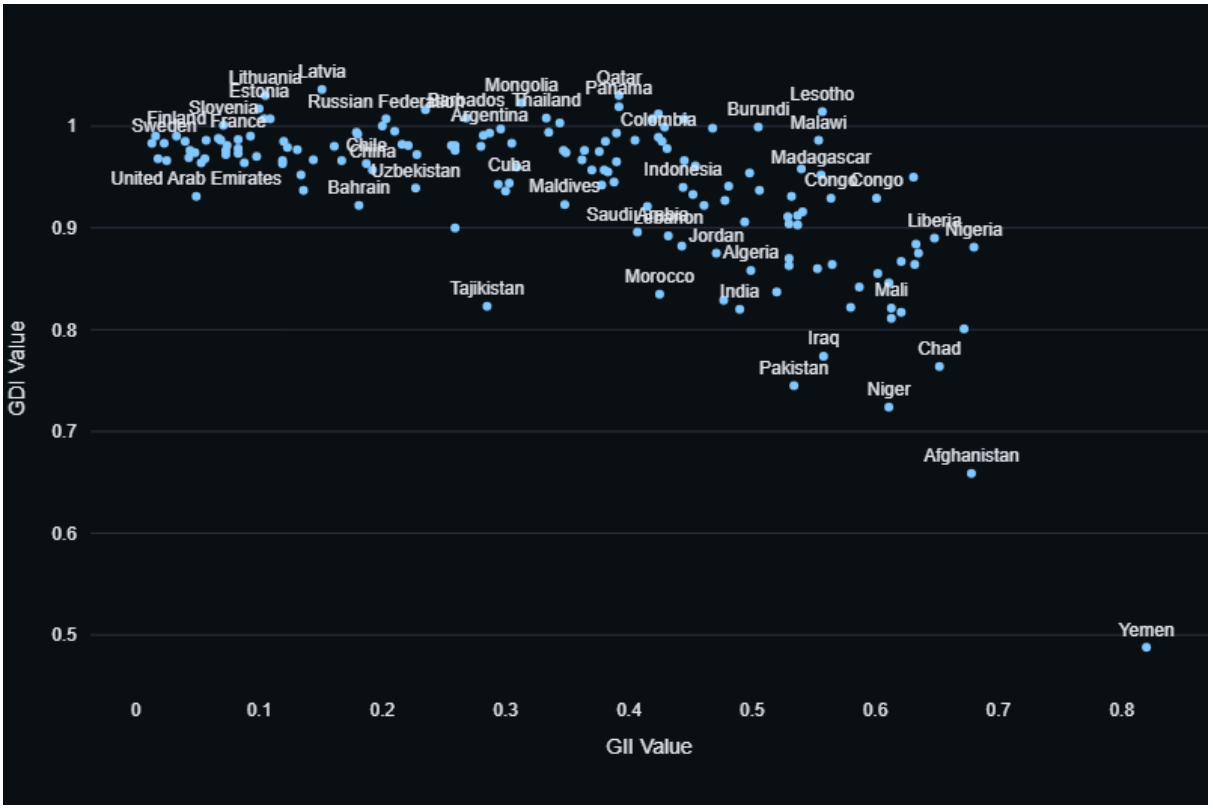


L'histogramme ci-dessus illustre les moyennes des trois catégories—santé reproductive, éducation, économie—qui composent le calcul du GII pour nos cinq clusters. En observant les moyennes, on remarque évidemment que le cluster 1 présente les valeurs les plus élevées, tandis que les clusters 2 et 3 semblent avoir des profils similaires, ce qui soulève la question de la pertinence d'avoir divisé les pays en cinq clusters au lieu de quatre.

En outre, le facteur de la santé reproductive se distingue comme le plus discriminant dans la répartition des pays par cluster. Ce constat souligne l'importance cruciale de la santé reproductive dans la classification des pays selon leur GII et suggère que c'est sur cet aspect que l'attention doit se concentrer pour réduire les inégalités de genre à l'échelle mondiale. En effet, les variations significatives dans ce facteur à travers les clusters suggèrent que des améliorations dans ce domaine, comme l'accès aux femmes à l'avortement ou à la contraception, peuvent potentiellement entraîner des progrès importants vers l'égalité des sexes.

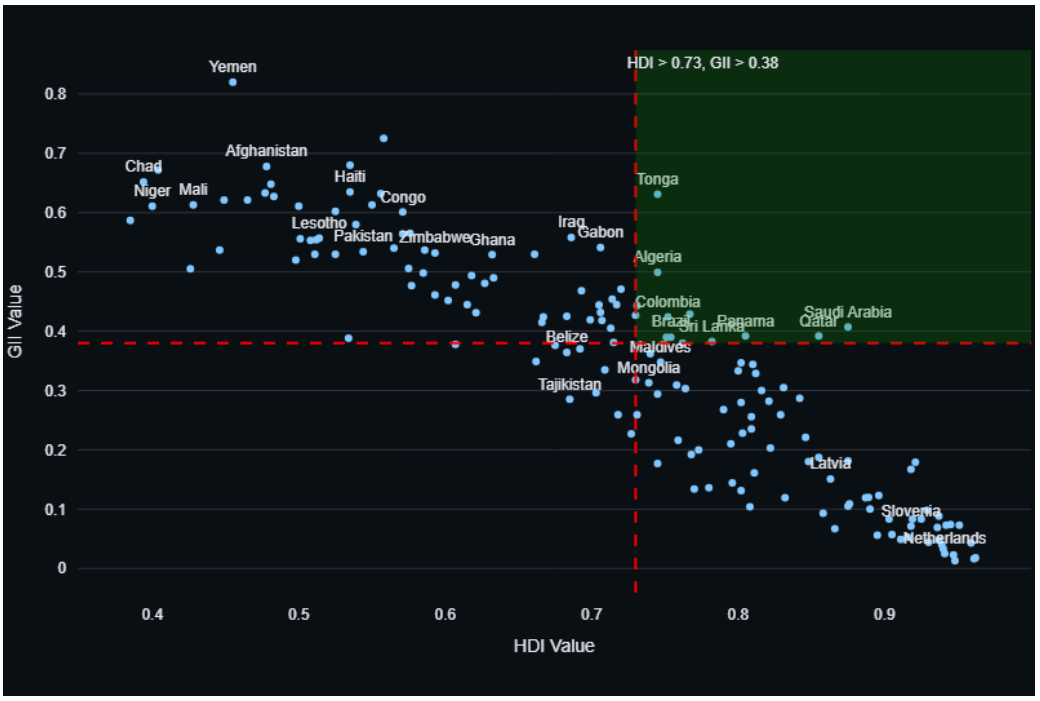
Comparons maintenant nos deux indicateurs afin de mettre en valeur leurs biais, et mieux comprendre la classification des pays.

Sur le graphique ci-dessous, j’ai affiché le nuage de point des pays avec, en abscisse, leur GII, et en ordonnée, leur GDI.



On peut faire plusieurs observations. Des pays tels que la Chine, la Moldavie et le Vietnam montrent un meilleur classement sur le GII par rapport au GDI, gagnant jusqu'à 40 places. Cela peut être attribué à des facteurs tels qu'un faible taux de mortalité maternelle et de fécondité chez les adolescentes, ainsi qu'un niveau élevé de participation des femmes au marché du travail. À l'inverse, l'Arabie Saoudite et le Qatar enregistrent une perte d'environ 50 places dans le classement du GII par rapport au GDI, due principalement à un faible taux d'emploi féminin et à une très faible représentation des femmes dans le gouvernement.
Ces variations de classement au sein des différents groupes de pays soulignent le problème du Gender Development Index dans la prise en compte du Revenu National Brut comme critère, avantageant les pays plus riches et occultant ainsi les discriminations à l’encontre des femmes dans ces régions développées.
Cette comparaison met en évidence l'importance de considérer différents indices pour une image complète de l'inégalité de genre et suggère que les politiques visant à réduire ces inégalités doivent être multifacettes et ciblées en fonction des problématiques spécifiques identifiées par ces indices.

Finalement, j'exprimais précédemment l'idée que les inégalités de genre paraissent largement corrélées au développement économique global du pays considéré, mais on peut remarquer quelques exceptions.



Certains pays, bien qu'ayant un HDI considéré comme "haut" par l'ONU (>0.7), possèdent néanmoins un GII supérieur à 0.38. Il s'agit des pays présents dans la partie que j'ai surlignée en vert sur le graphique; on retrouve par exemple l'Arabie Saoudite et le Qatar comme cité précédemment, mais également le Brésil, l'Algérie, la Colombie, le Panama…

Conclusion

Finalement, les différents affichages permis par la plateforme donnent une image globale plutôt représentative des inégalités de genre dans le monde, et permet également d'avoir un avis critique sur les différents indicateurs qui les mesurent.
Cela dit, la plateforme n'est pour l'instant adaptée qu'au traitement de mes deux jeux de données spécifiques, et l'onglet prédiction n'est que très peu concluant dans son état actuel.

Globalement, les inégalités de genre sont encore trop dépendantes de la situation économique et politique des régions du monde pour réussir à mesurer la fiabilité d'éventuelles prédictions, et les indicateurs actuels, qui ne prennent pas en compte des facteurs comme les violences domestiques et sexuelles, le taux de propriété ou la participation à la communauté, sont encore trop éloignés de la réalité.