Projet info S7

Septembre :

Début du projet : lecture du livre, apprentissage du fonctionnement de Streamlit

Octobre :

Création de la plateforme Streamlit en s’inspirant de celle du projet 1A statistique.

Récupération du dataset de Fatine (GII), et réajustement de la plateforme streamlit pour qu’elle puisse le lire.

Novembre :

Recherche de nouveaux datasets (plateformes Kaggle, data.gouv etc…) mais difficultés à trouver des jeux de données intéressants : finalement ceux que j’utilise proviennent du site Human Development Reports (<https://hdr.undp.org/data-center/thematic-composite-indices>) et concernent le GDI et le GII. Ainsi que ceux de la plateforme kaggle (accès aux médicaments, passeport etc…)

Création de classification en groupe de pays.

Difficultés : normalisation des données. Elles n’ont pas toutes les mêmes noms de variables, les mêmes unités… (1 ou 0, variables qualitatives, pourcentage… etc) -> décision de se focus seulement sur qq jeux de données, et la plateforme devra donc avoir un cas différent en fonction de chaque dataset

Perspective : affichage de cartes, + apprentissage automatique sur les jeux de données qui contiennent des infos sur les 20 dernières années, afin de faire des prédictions (cartes d’évolution ?)

Décembre :

Affichage sur une même plateforme des cartes du GII, et du GDI en fonction des années.

Difficultés : trouver un fichier GeoJSON qui ait exactement les mêmes noms de pays que mes datasets pour que tous soient reconnus (pour l’instant la Russie ne peut pas être coloriée car elle s’appelle « Russian Federation » dans mes datasets, et « Russia » sur mon fichier GeoJSON)

Peut-etre chercher sur Natural Earth ? <https://www.naturalearthdata.com/downloads/>

Idée : plateforme qui affiche plutôt une carte du monde vierge, qui se colorie ensuite en fonction d’un aspect qu’on selectionne (GDI, GII, année, yes or no infos…) + superposition des cartes ? Affichage des groupes que j’ai fait sur la carte

Retouches : la légende pour genderpassport (il faudrait juste oui ou non, rouge et vert pas le dégradé)

Pour les oui/non : faire la liste des pays ou =NON en (year)

Elle veut des graphiques en 3D ?

Apprentissage auto :

Page prédiction test\_machinelearn.py

Fichiers geoJSON : chatgpt m’a donné deux dépôts git ou il ya des fichiers geojson de cartes du mondes qui reconnaissent les pays avec leurs codes ISO3. Ce serait intéressant d’essayer de les matcher avec la carte. Chnger les couleurs en fonction de si on a pas de données ou si juste c pas le bon nom, comme àa je pourrais changer les noms à la main

Une corrélation avec Pay\_gap\_europe  et Human\_development?

18 décembre :

Apprentissage auto : méthode Ridge, linear regression avec valid croisée, KNN, SVR et Random forest à retravailler (ajuster les paramètres car pour le moment l’affichage des graphiques ca donne que des droites c pas du tout précis) COMPRENDRE LES MODELES QUE J’UTILISE

Modele ARIMA qui s’adapte bien aux séries temporelles et donne un modèle assez réaliste si on regarde sur les 10 prochaines années : essayer de joindre les deux bouts de courbe (FAIT) Comment on ajuste les paramètres du modele ? (5,1,0)

Animation de graphiques ?

Colorier ma carte avec les valeurs prédites grâce au modèle ARIMA + colorier les pays selon leurs groupe

Soutenance semaine du 29 janvier ?

https://www.datahub.io/core/geo-countries#python

<https://github.com/datasets/geo-countries/blob/master/datapackage.json> (geo json donné par chatgpt)

<https://raw.githubusercontent.com/python-visualization/folium/master/examples/data/world-countries.json>

https://hdr.undp.org/data-center/thematic-composite-indices/gender-inequality-index?gad\_source=1&gclid=CjwKCAiAx\_GqBhBQEiwAlDNAZo4F51k0R1X9JYp78P5pX\_gDPa6dEk1h7DEdS5eWA1kC0amSDD0xzxoCA0IQAvD\_BwE#/indicies/GII

<https://hdr.undp.org/gender-development-index#/indicies/GDI>

**ce que j’essaye de faire :**

* Une fois que ça marche : colorier l’histogramme en 4 couleurs, 1 neutre puis 3 genre jaune rouge vert pour Santé, Education et Economie

Pour « Gender\_Inequality\_Index » :

Santé : « Maternal\_mortality » « Adolescent\_birth\_rate »

Education : « F\_secondary\_educ » « M\_secondary\_educ »

Economie : « F\_labour\_force » « M\_labour\_force »

Pour « GDI\_detail\_2019 » :

Santé : « Lif\_Expec\_Female », « Lif\_Expec\_Male » «

Education : « Excep\_Yrs\_Schooling\_Female » « Excep\_Yrs\_Schooling\_Male » «Mean\_Yrs\_Schooling\_Female » «Mean\_Yrs\_Schooling\_Male »

Economie : « GNI\_PC\_Female » « GNI\_PC\_Male »

* Faire sur un graphique pour un pays de chaque groupe, la participation de chaque facteur (santé education autonomie) au GII comme sur le rapport

Regrouper ma plateforme principale selon ce que je veux comme orga sur mes slides

Corrélation : mettre que la valeur par défaut de la selectbox 2 soit autre chose que celle de la première select\_box

La corrélation (positive) et négative n’est pas bonne, ça donne tout le temps « positive »