Rapport de projet 2A :

A quoi ressemblerait le monde si les données relatives aux minorités étaient prises en compte ?

Ecole des Mines de Nancy

Réalisé par : Fatine RHANEM

Encadré par : Dominique Benmouffek

Année Scolaire : 2022/2023

Remerciements

Je souhaite tout d'abord adresser mes remerciements au corps professoral et administratif de l'Ecole des Mines de Nancy , pour la qualité de l'enseignement et pour l'opportunité de réaliser ce projet. A cette occasion, je veux exprimer ma sincère gratitude à Madame Dominique Benmouffek pour sa confiance, sa disponibilité tout au long de la réalisation du projet et sa pertinence quant aux décisions relatives au projet.

Résumé

Dans le cadre du projet 2A en deuxième année en département informatique, j'ai travaillé sur le sujet : « A quoi ressemblerait le monde si les données relatives aux minorités étaient prises en compte ? » , qui porte sur l'analyse de données en suivant plusieurs étapes cruciales commençant par la collecte, le traitement ainsi que l'utilisation des données pour la prise de décision. Le choix pertinent d'un jeu de données peut changer la teneur de leur interprétation dans un sens comme dans l'autre.

A travers ce travail, j'ai essayé de mettre en évidence les biais qui existent dans les données et les algorithmes faits pour les traiter et les analyser. Par conséquent, monter que les groupes minoritaires souffrent d'une certaine discrimination qui fait qu'ils ne soient pas présents et inclus dans la prise de décision.

Ce projet m'a permis de développer des compétences techniques en utilisant des outils tels que Python et Streamlit pour traiter, analyser et visualiser les données. Pour ce faire, j'ai procédé à la création de deux modèles de traitement de données, extraites de Kaggle, pour mettre en relief les différences de revenus selon plusieurs facteurs et la variation de l'index des inégalités de genre. Ce travail avait pour but de montrer l'importance de prendre en compte les minorités afin de prendre des décisions justes et équitables envers eux.

Abstract

As part of the 2A project in the second year in the Mining school of Nancy in the computer science department, I worked on the subject : “ What would the world look like if data related to minorities were taken into account? “ . It focuses on data analysis by following several crucial steps starting with the collection, processing and use of data for decision-making, which will have an impact on thousands of people around the world.

Through this work, I tried to highlight the biases that exist in the data and the algorithms made to process and analyze them. Therefore, to show that minority groups suffer from a certain discrimination that makes them not present and not included in decision-making.

This project allowed me to develop technical skills using tools such as Python and Streamlit to process, analyze and visualize data. To do this, I created two data processing models, extracted from Kaggle, to highlight the differences in income according to several factors and the variation of the gender inequality index.

This work aimed to show the importance of taking minorities into account in order to make fair and equitable decisions towards them.

Liste des figures :

Table des matières :

Introduction générale

Le sujet « A quoi ressemblerait le monde si les données relatives aux minorités étaient prises en compte ? » soulève un questionnement très profond dans le domaine de l'intelligence artificielle et l'analyse de données à propos des minorités qui sont souvent ignorées en amont dans le choix des données de référence.

En effet, les jeux de données, utilisées pour entraîner et valider des modèles de « Machine Learning », sont souvent biaisées et peuvent présenter des erreurs en défaveur des groupes de minorités. Par conséquent cela peut engendrer des discriminations dans les décisions prises à partir de ces données.

C'est pourquoi ce problème est pertinent et nécessite d'être étudié puisqu'on vit dans un monde où les décisions sont prises de manière automatisées dans de nombreux domaines. Il est donc primordial de prendre des décisions équitables et justes, sans prendre en considération la race, la religion, le genre, la couleur ainsi que d'autres. Pour arriver à cet idéal, il faut encourager l'inclusion des minorités qui sont les grands oubliés de certaines sphères de vie en collectant les données représentatives de la diversité, en les traitant et en développant des algorithmes. Ce long processus est crucial et doit prendre en compte les différences entre les groupes majoritaires et les groupes minoritaires.

C'est dans ce contexte que cette question est soulevée pour faire une analyse approfondie et une réflexion critique puisque ce sujet impactera des millions de personnes à travers le monde et pourra voir des conséquences à long terme surtout sur les minorités .

Chapitre 1 :

Comment prendre en compte les données relatives aux minorités ?

Introduction :

La collecte de données inclusives est essentielle pour garantir une représentation précise et équilibrée de la diversité des populations, y compris des minorités.

I. L'importance de la collecte de données inclusives :

Voici quelques éléments importants à considérer :

- Sensibilisation et engagement : Il est important de sensibiliser les collecteurs de données à l'importance de la collecte inclusive et de les encourager à recueillir des données représentatives de toutes les catégories de minorités.
- Participation active des minorités : Il est nécessaire d'impliquer activement les minorités dans le processus de collecte des données. Cela peut être fait en leur donnant une voix dans la définition des objectifs de la collecte de données, en leur offrant des incitations à participer et en garantissant la confidentialité et la sécurité des informations collectées.
- Diversité des sources des données : Il est important de diversifier les sources de données utilisées, en incluant des données provenant de différentes régions géographiques, de différents groupes socio-économiques et de différentes communautés culturelles. Cela permet d'éviter les biais potentiels liés à la concentration des données sur certaines populations.

II. Les méthodes de traitement des données pour prendre en compte les minorités :

Une fois que les données inclusives ont été collectées, il est nécessaire de les traiter de manière approprié pour tenir en compte des minorités. Voici quelques méthodes utilisées :

- Normalisation des données : La normalisation est essentielle pour éviter les biais indésirables. Cela implique de pondérer les données en fonction de la représentativité de chaque groupe, afin de garantir une évaluation des caractéristiques et des performances.
- Analyse des biais : L'analyse consiste à évaluer et à quantifier les biais présents dans les données. Cela peut être réalisé en comparant les distributions de caractéristiques et de résultats entre les différents groupes.
- Prétraitement des données : C'est une étape cruciale pour réduire les biais potentiels. Cela peut inclure la suppression des valeurs aberrantes. Le

prétraitement peut aider à rendre les données plus représentatives de la diversité de la population.

III. **Les outils informatiques pour aider à la prise en compte des données relatives aux minorités :**

Voici quelques exemples d'outils et de technique utilisées pour le traitement des données.

- Bibliothèques d'apprentissage automatique : Plusieurs bibliothèques et frameworks ont été développés pour faciliter l'application de méthodes d'apprentissage. Par exemple Streamlit et le langage Python.
- Techniques de prétraitement des données : Les outils informatiques offrent beaucoup de fonctionnalités pour effectuer des opérations de prétraitement. Des bibliothèques comme Scikit-learn en Python fournissent des outils puissants pour réaliser ces opérations de manière efficace.
- Visualisation des données : Les outils de visualisation des données jouent un rôle dans l'interprétation des résultats liés aux minorités. Des bibliothèques telles que Matplotlib et Ploty permettent de créer des graphiques et des visualisations interactives pour mettre en relief les inégalités et les disparités entre les groupes.

Conclusion :

En prenant en compte les données relatives aux minorités, la prise des décisions est plus juste et plus éclairée. Cela nécessite un engagement collectif pour le collecte des données et l'utilisation de techniques appropriées pour garantir une représentation précise de la diversité de la population.

Chapitre 2 :

Mise en pratique : création de deux modèles de traitement de données avec Python et Streamlit

Introduction :

L'objectif de cette mise en pratique est d'utiliser Python et Streamlit pour explorer et visualiser les données, appliquer des techniques de prétraitement des données et créer deux modèles de traitement des données qui tiennent compte des minorités.

I. Premier modèle :

1. Présentation du jeu de donnée utilisé :

Le jeu de données utilisé s'intitule ' Adult Census Income '. Il contient 32 561 entrées qui représentent chacune une personne, 14 variables dont quelques-unes sont des attributs individuels tels que l'âge, le niveau d'éducation, le sexe, le statut marital, la relation familiale, la race, l'occupation et d'autres. La variable cible est 'income', qui précise si le revenu est supérieur à 50000 dollars ou non. Par conséquent, ce jeu de données permet d'explorer les disparités de salaires en fonction de plusieurs facteurs influant.

2. Création du modèle de traitement de données avec Python et Streamlit :

Chargement des données : on charge le jeu de données à partir du fichier CSV « adult.csv ».

Prétraitement des données : Avant de procéder à l'entraînement du modèle , on doit séparer les variables d'entrée (X) et la variable de sortie ou cible (Y). Les variables catégorielles sont encodées en utilisant la technique de One-Hot Encoding pour les transformer en variables numériques prêtes à être entraîner.

Entraînement du modèle : J'ai utilisé un modèle de régression logistique. On entraîne le modèle sur les données d'entraînement (X_train , Y_train).

Le but de ce code est de permettre à l'utilisateur d'entrer des valeurs pour les différents attributs via une interface Streamlit. Le résultat de prédiction qui est le revenu de la personne indique si celui-ci est supérieur ou inférieur à 50K $ annuel.

3. Interprétation des résultats :

On remarque que le revenu varie selon plusieurs facteurs, ce qui peut nous aider à savoir l'impact sur les minorités. Le modèle de régression logistique permet d'identifier les relations entre les variables et le revenu.

En observant les coefficients du modèle, on peut extraire les attributs les plus importants. Par exemple le niveau d'éducation, l'occupation et le pays d'origine qui jouent un rôle fondamental dans la détermination du revenu.

- Identification des groupes minoritaires :

En examinant les disparités importantes de revenu en ce qui concerne les groupes qui sont représentés dans ce jeu de données, on peut identifier les groupes minoritaires représentés par les variables ‘race’ et ‘sexe’.

Il est important de noter que certains groupes minoritaires sont en mesure d’obtenir des revenus inférieurs à 50 000 $ par rapport à d’autres groupes . Cette analyse met en évidence le fait que la disparité significative est due à plusieurs facteurs tels que les différences d’accès à l’éducation, l’occupation, les heures de travail par semaine ce qui les privent d’être mieux représentés dans des domaines professionnels bien rémunérés.

- Analyse des distributions de revenu :

Il est important de déterminer les distributions de revenu pour chaque groupe minoritaire selon la race :

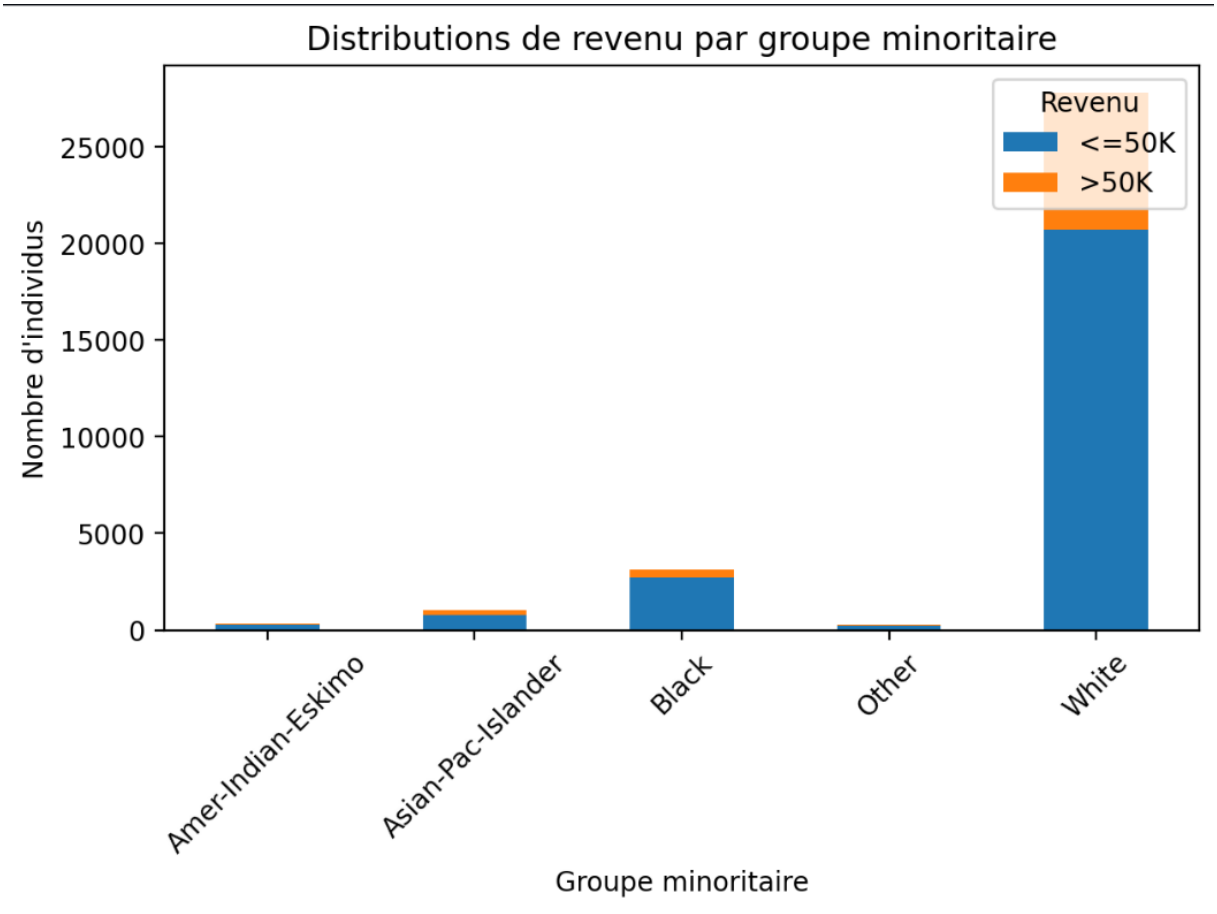

Figure 1 : Distribution de revenu par groupe de race

D’après la figure ci-dessus, on remarque que le revenu est inférieur à 50K pour la plupart des blancs (groupe majoritaire). En ce qui concerne les groupes minoritaires, dont les noirs constituent la plus grande proportion, on observe que le revenu est généralement inférieur à 50K $.

Les distributions de revenu pour chaque groupe minoritaire selon le sexe :

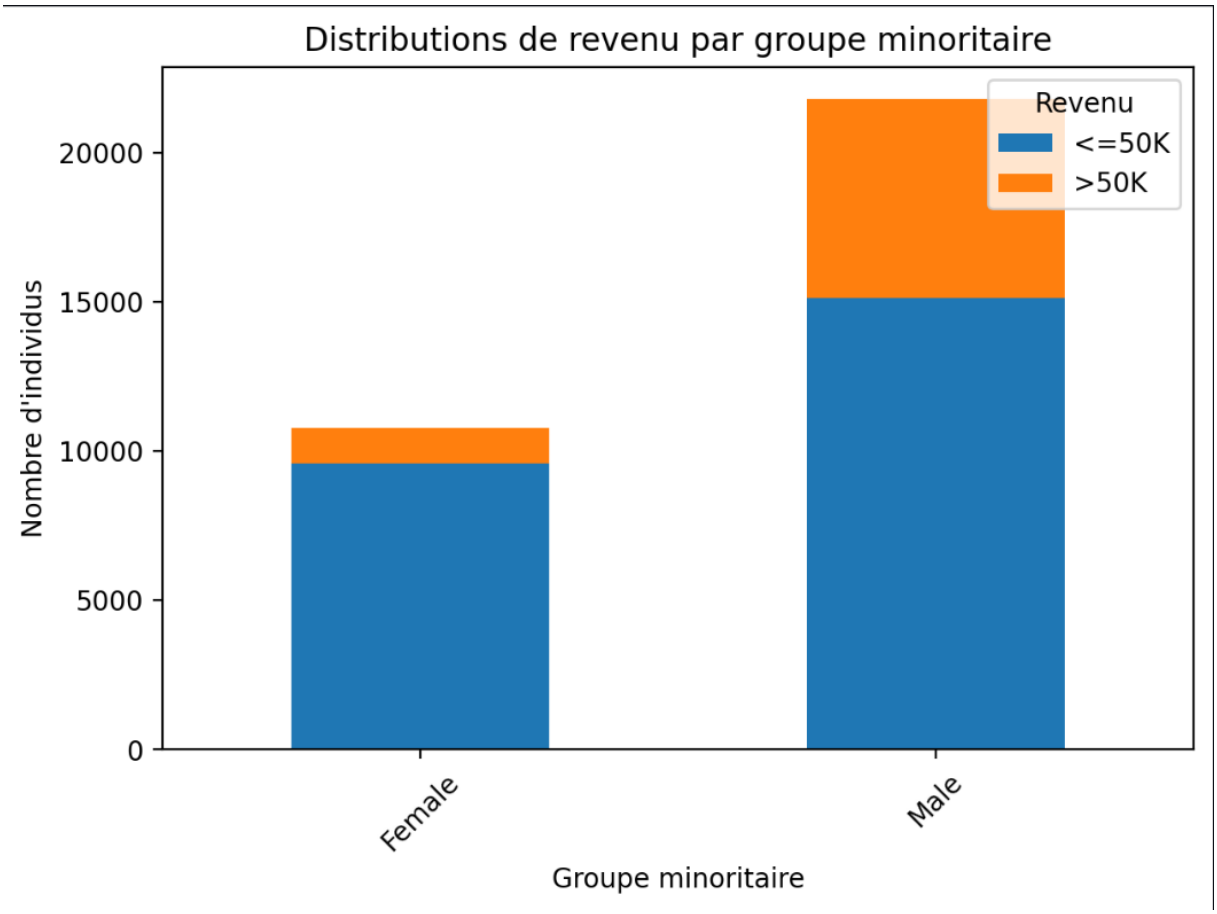

Figure 2 : Distribution de revenu par groupe de sexe

D'après la figure ci-dessus, on remarque que les femmes sont moins représentées que les hommes, ce qui pourrait signifier que les femmes ne travaillent pas en majorité, ou ne sont pas rémunérées. On peut conclure que les hommes sont plus susceptibles d'avoir un revenu supérieur à 50K $.

- Analyse des variables explicatives :

  Le but est d'identifier les variables les plus significatives. Pour faire ceci, il est nécessaire d'adopter des méthodes telles que la régression linéaire .
  On remarque que le coefficient de la variable 'capital-gain ' a un coefficient de 0.000338 ce qui signifie que le revenu et cette variable sont positivement corrélés .(une augmentation du capital-gain est associée à une augmentation de revenu).
  La variable 'age' a un coefficient -0.007313 ce qui veut dire qu'elle est négativement corrélée au revenu 'income' (une augmentation de l'âge introduit une diminution du revenu).

  ⇨ Les variables dont les coefficients sont positifs associent une augmentation du revenus, tandis que les variables dont les coefficients sont négatifs introduisent une diminution du revenus.

    - Interprétation des résultats :

D’après les résultats obtenus, on peut remarquer que l’âge, le niveau d’éducation, le pays d’origine et l’occupation ont un impact significatif sur les revenus des groupes minoritaires.

II. Deuxième modèle :

1. Présentation du jeu de données utilisé :

Le jeu de données s’intitule ’Gender Inequality Index ‘ et fournit l’indice d’inégalité entre les genres dans différents pays. Il examine la disparité entre les hommes et les femmes. Il comprend 11 variables telles que le pays, l’année, la part des sièges qui sont occupés par les femmes au parlement, le taux de participation à la force de travail par les femmes et les hommes. Par conséquent ce jeu de donnée permet de se concentrer sur les domaines où l’inégalité est plus prononcée entre les femmes et les hommes, donc encourager à mettre plus d’efforts pour promouvoir l’égalité .

2. Création du modèle de traitement de données avec Python et Streamlit :

Les étapes de la création du modèle :

- **Affichage du jeu de données** : Visualisation des données brutes ‘gender_inequality_index.csv’.
- **Description du jeu de données** : Fournir des informations statistiques telles que : la moyenne, l’écart-type , les valeurs maximales et minimales.
- **Matrice de corrélation** : Indiquer les relations linéaires entre les variables .

matrice:

	GII	Rank	Maternal_mortality	Adolescent_birth_rate	Seats_parliament	F_sec
GII	1.0000	0.9968	0.7135	0.8068	-0.4241	
Rank	0.9968	1.0000	0.7336	0.8208	-0.4199	
Maternal_mortality	0.7135	0.7336	1.0000	0.7528	-0.1622	
Adolescent_birth_rate	0.8068	0.8208	0.7528	1.0000	-0.0947	
Seats_parliament	-0.4241	-0.4199	-0.1622	-0.0947	1.0000	
F_secondary_educ	-0.8093	-0.8117	-0.6980	-0.7287	0.1695	
M_secondary_educ	-0.7821	-0.7816	-0.6426	-0.6918	0.1688	
F_Labour_force	-0.0710	-0.0508	0.2305	0.2605	0.2790	
M_Labour_force	0.1583	0.1601	0.1062	0.2639	0.0597	

Figure 3 : Matrice de corrélation

D’après la figure ci-dessus , on remarque que le GII (gender inequality index) est fortement corrélé à : Rank, ‘Adolescent_birth_Rate’ , ‘F_secondary_educ’ .

- **Séparation des données** : Diviser le jeu de données en ensemble d’entraînement (X_train , Y_train) et de test (X_test , Y_test) .
- **Entraînement du modèle** : Utilisation d’un modèle de régression linéaire afin de prédire l’indice d’inégalité des genres en fonction des autres attributs.

3. **Interprétation des résultats :**
    - **Histogramme de GII :** représenter la distribution de la variable GII du jeu de données.

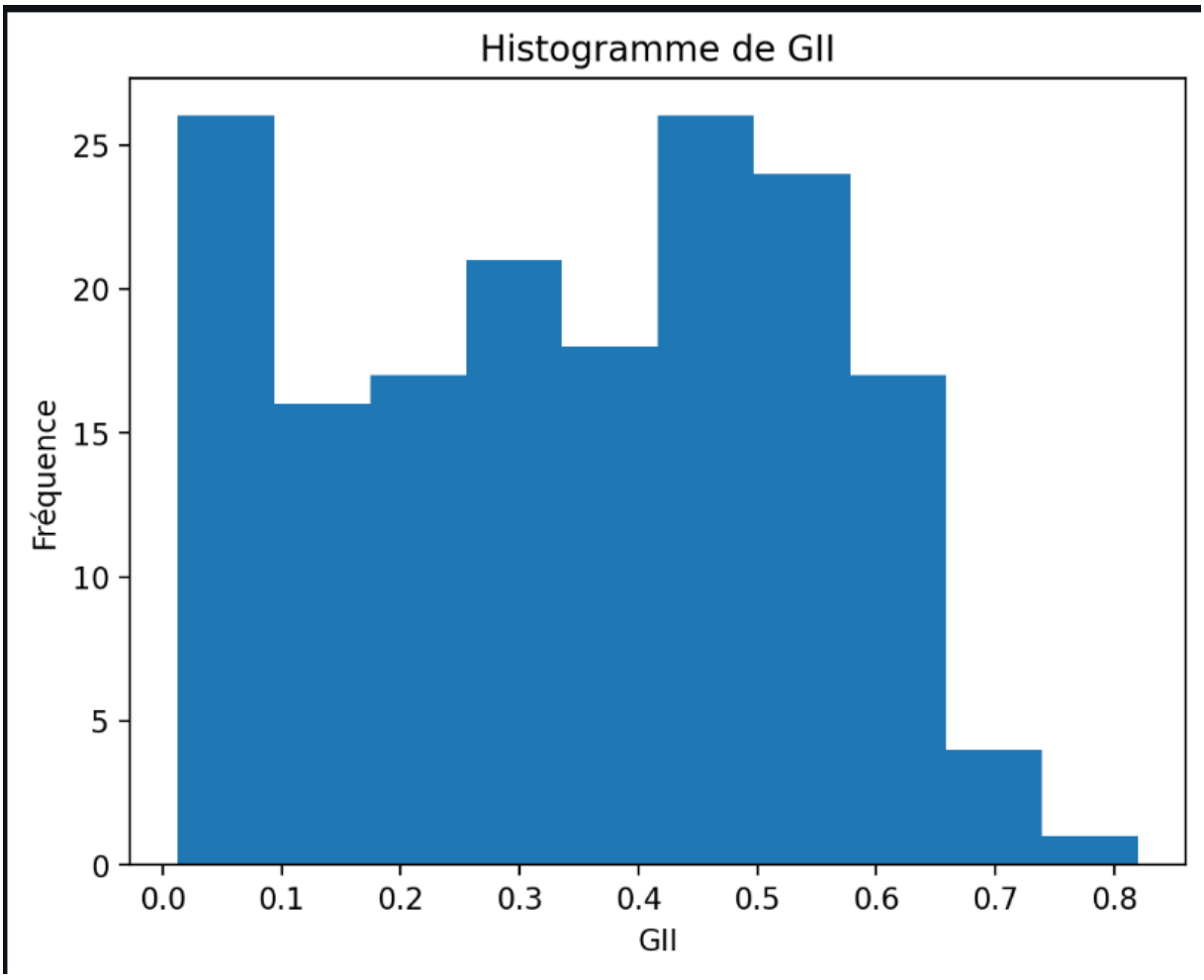

Figure 4 : Histogramme de GII

D’après la figure ci-dessus, on remarque que le GII le plus fréquent est celui dont la valeur est comprise entre 0.02 et 0.1 et entre 0.42 et 0.5. Au contraire, le GII le moins fréquent est celui dont la valeur est comprise entre 0.73 et 0.81.

- **Diagrammes en barres :** Le but est d’exprimer le GII en fonction de chaque valeur de la variable choisie.
  Selon le développement humain (Human_development) :

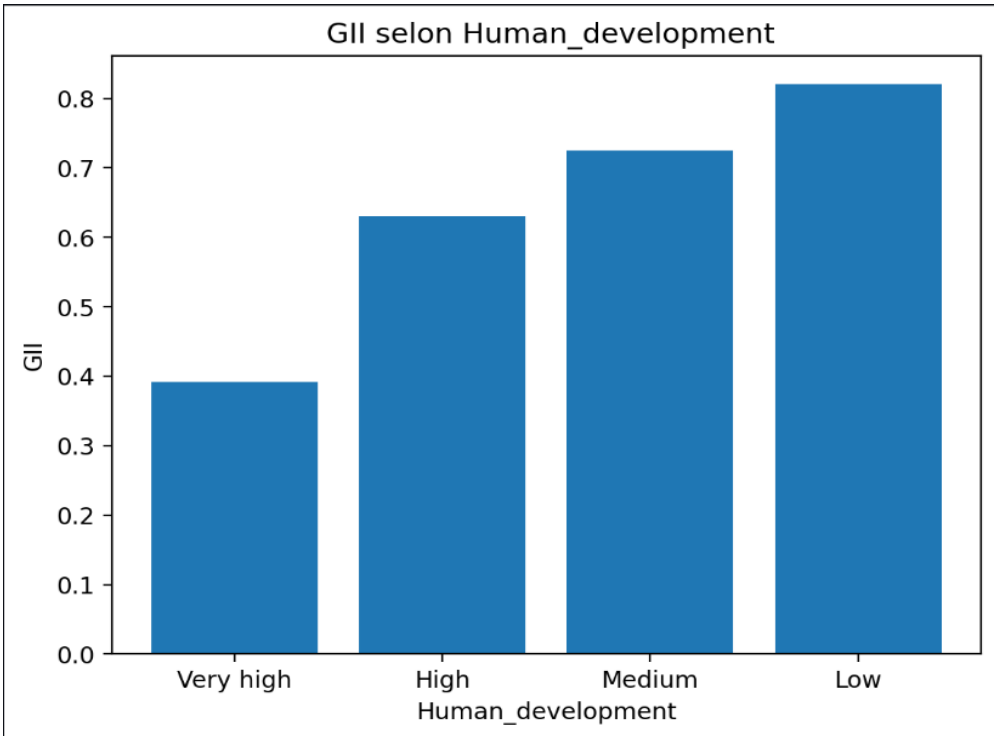

Figure 5 : GII selon la variable Human_development

On remarque que les pays dont le développement est faible présente un GII supérieur, dont la valeur maximale est de 0,82. Par contre, les pays très développés présentent un GII inférieur à 0,4.

Selon les places dédiées aux femmes au parlement (Seats_parliament) :

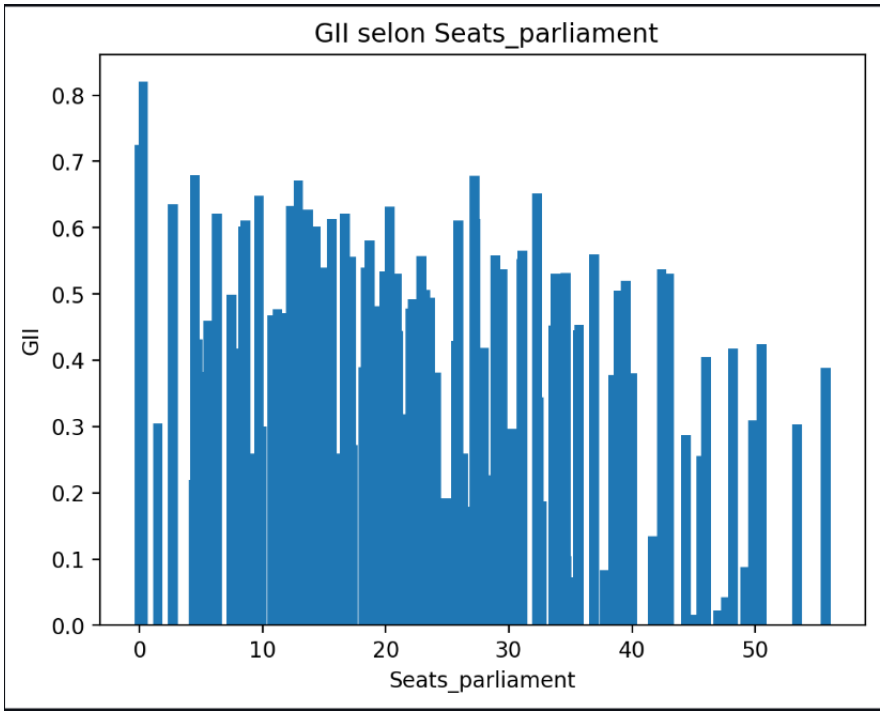

Figure 6 : GII selon la variable Seats_parliament

D’après la figure ci-dessus, on remarque que le GII diminue dès que la valeur de la variable augmente et vice versa.

Conclusion :

Ce chapitre a présenté la mise en pratique de la création de deux modèles de traitement de données avec Python et Streamlit, en utilisant les jeux de données déjà présentés. Les différentes étapes du processus, notamment le chargement du jeu de données, l’exploration, le prétraitement des données et la création des modèles, ont été abordées.

Conclusion générale :

Ce travail a été fait dans le cadre du projet en deuxième année au département informatique à l'école des Mines de Nancy.

A l'issue de ce projet, j'ai souligné l'importance de collecter des données inclusives et d'utiliser des techniques appropriées pour réduire les biais et garantir une représentation précise des minorités.

Au niveau de ce projet, j'ai examiné les outils informatiques qui peuvent faciliter la prise en compte des données relatives aux minorités. Ces outils, tels que les bibliothèques de prétraitement des données et les bibliothèques d'apprentissage automatique, offrent des fonctionnalités essentielles pour soutenir les pratiques équitables de traitement des données.

En conclusion, en prenant en compte les minorités, nous pouvons contribuer à atténuer les inégalités et à promouvoir une société plus juste et équitable. Cela nécessite un engagement continu et collectif pour mettre en pratique les concepts discutés dans ce projet et pour faire évoluer nos pratiques de traitement des données vers plus d'inclusion et d'équité.