IFT3700 A2022 Devoir 2

(Version préliminaire)

Directives

- Remise électronique avant le 22 décembre 23h55.
- Le travail doit être fait en équipe de 3.
- Le travail peut-être rédigé en français, anglais, espagnol, portuguais ou italien.

Description

Vous allez devoir collecter et nettoyer des données numériques et catégoriques au sujet des pays du monde. Les données proviennent de 40 tableaux disponibles sur Wikipédia. En plus de traiter les valeurs numériques données, toutes les caractéristiques (colonne), seront discrétisées.

Vous devrez remettre un rapport en format pdf, le code source python, et un certain nombre de fichiers en format json. La qualité de la présentation du rapport est importante. Il est important de commenter sommairement les résultats obtenus. La qualité de votre analyse ne sera pas à juger en fonction de la connaissance géopolitique qui ne serait pas présente dans les données que vous avez à analyser.

Il est important de se rappeler que les données observationnelles ne permettent pas seules d'établir des liens causaux, il est impératif que votre analyse ne contienne pas de jugement de valeur.

1. **Collecte de données**
    a. Collecter les 40 colonnes telles que spécifiées dans la section donnée.
    b. Il est recommandé que pour chacune des colonnes calculer la moyenne, la médiane, le maximum, le mode, le minimum, la variance, le nombre de valeurs manquantes. Accès rapide et facile à

ses valeurs vous aidera à faire une analyse compétente des résultats.

c. Éliminez les pays ayant 12 ou plus valeurs manquantes. Il est probable que près de la moitié des pays et territoires seront ainsi éliminés.

d. Premièrement, faire une étape 0 où les valeurs manquantes sont remplacées par la médiane. Utiliser ensuite une régression linéaire pour obtenir une meilleure évaluation de chacune des données manquantes. Effectuer le même processus avec la 1re régression linéaire pour faire une 2e régression linéaire de meilleure qualité qui sera les résultats que vous allez utiliser.

e. Dupliquer chacune des 40 colonnes. Pour chacune des colonnes dupliquées, remplacer les valeurs inférieures (ou égales) à la médiane par zéro et celle supérieure à la médiane par un.

2. **Corrélations**
   La corrélation est toujours fonction des variables continues. Lorsque l'on parle de la force d'une corrélation, il s'agit de la valeur absolue.

   a. Vous devez calculer le coefficient de corrélation entre chaque paire de colonnes ayant une valeur numérique.

   b. Pour chaque colonne, identifier la colonne pour laquelle la corrélation est la plus forte.
   Remise: liste de “int” (json)

   c. Ordonner, en ordre décroissant, les colonnes par leur corrélation absolue moyenne avec les autres. Cet ordre sera utilisé à la question 2.5.a
   Remise: liste de “int” (json)

3. **Prédictions avec classifieur bayésien et régression linéaire**
   Pour toutes les questions qui suivent, vous devez utiliser les deux techniques pour effectuer des prédictions. La régression linéaire se fait avec les valeurs réelles et le classifieur bayésien avec les valeurs binariser. Il n'est pas nécessaire de normaliser les valeurs pour une régression linéaire et même il est préférable souvent de ne pas le faire au niveau de l'interprétabilité, lorsqu'il s'agit d'évaluer l'impact moyen il est clair

dans ce cas que l'échelle a de l'importance et je vous recommande donc de normaliser les valeurs comme une côte z.

a. Pour chaque colonne, analyser la qualité des prédicteurs en fonction des autres colonnes.

b. Pour chaque colonne, identifier la paire de colonnes qui permet de la prédire avec la meilleure précision.
Remise: liste de listes json

c. Trouver les deux colonnes qui nous donnent le meilleur prédicteur pour chacune des autres caractéristiques (en moyenne).

4. Visualisation et représentations

a. Vous devez appliquer un algorithme de réduction de dimensionnalité pour pouvoir afficher en 2 dimensions l'ensemble des pays. Vous êtes libre d'utiliser la technique qui vous semble la plus appropriée. Vous devez produire une image en 2 dimensions où on peut observer la position de chaque pays. La création d'une image de grand format vous permettra d'afficher à côté de chaque point le nom du pays même si celui-ci est relativement petit dans l'image.

b. Analyser le résultat d'une réduction sur 2 dimensions et d'une réduction sur 5 dimensions en essayant de trouver une interprétation aux valeurs de chacune des dimensions réduites.

Données

La liste suivante contient une référence précise à la page, au tableau, à la colonne pour chaque caractéristique a étudier. Notez qu'il existe des outils qui permettent de faciliter le téléchargement de ce genre de tables.

Lorsque les valeurs sont données sous forme d'intervalle, vous devez choisir le point médian entre la limite inférieure et la limite supérieure de l'intervalle.

Dans les très rares cas où deux valeurs sont données, c'est-à-dire lorsque différentes sources différentes valeurs, vous devez choisir le point médian.

1. List of countries by GDP (nominal) per capita, GDP (in US$) per capita by country, United Nations, estimate.
2. List of countries by Internet connection speeds, Fixed broadband, Average download speed (Mbit/s) (Ookla).
3. List of countries by alcohol consumption per capita**,** Recorded per capita consumption of pure alcohol (litres) per adult 15 years of age and over per year, 2016.
4. List of countries by intentional homicide rate, Intentional homicide victims per 100,000 inhabitants. From UNODC, rate.
5. List of countries by military expenditures, List by the Stockholm International Peace Research Institute, % of GDP.
6. List of countries by Human Development Index, 2019 data.
7. Democracy Index, By country, 2020.
8. List of countries by tertiary education attainment,Countries by level of tertiary education, at least a 2-year tertiary degree or its equivalent %.
9. Importance of religion by country, Countries/Districts, Yes important.
10. Christianity by country, UN members and dependent territories, % Chriastian.
11. Islam by country, Table, Muslim percentage of total population.
12. Buddhism by country, Buddhist population by country, % population.
13. Jewish population by country, Table, Core population, pct.
14. List of countries by infant and under-five mortality rates, Under-five mortality (deaths/1,000 live births) – 2019 estimates
15. Age of criminal responsibility, By country, age reduced.
16. List of countries by minimum wage, List, Annual, US$.
17. List of countries by external debt, List of countries with respect to external debt, % GDP

18. List of countries by income equality, UN, World Bank and CIA list – income ratios and Gini indices, WB Gini %.
19. List of countries by total health expenditure per capita, Total health expenditure per capita in 2018 PPP international U.S. dollars, 2018.
20. List of countries by suicide rate, Suicide rates by sex and country, All.
21. List of sovereign states and dependencies by total fertility rate, 2019 List by the World Bank, 2019.
22. Tobacco consumption by country, Prevalence of current tobacco use among persons aged 15 years and older, 2000.
23. List of countries by obesity rate, list, 2016 %.
24. List of countries by number of Internet users, Table, %.
25. List of countries by median age, Median age per CIA World Factbook 2018 and 2020 (ranked) estimates, 2020 median combined.
26. List of countries by economic freedom, 2019 Economic Freedom of the World Index, score.
27. List of countries by oil production, per capita 2017. **Valeur absente = 0, valeur manquante [–] = comme les autres.**
28. List of countries by population growth rate, Table, UN 2015-20.
29. List of countries by life expectancy, List by the World Health Organization (2019), Life expectancy at birth, All.
30. List of countries by meat consumption, Meat consumption by country, KG per capita **2002**.
31. List of countries by incarceration rate, Incarceration populations and rates. From World Prison Brief, rate per 100000.
32. List of countries by literacy rate, List of UN member states by age group and gender disparity, Elderly (65+), rate.
33. List of countries by age at first marriage, Woman. Les différents tableaux doivent être combinés.
34. List of countries by spending on education, % of GDP
35. List of countries by homeless population, per 100000.
36. List of countries by milk consumption per capita, 2013.
37. List of countries by number of scientific and technical journal articles, per capita.
38. Books published per country per year, titles (total).
39. List of countries by food energy intake, kilocalories.
40. List of countries by average yearly temperature, celsius.