MANAGEMENT & DATASCIENCE

SOBRIÉTÉ NUMÉRIQUE : DÉVELOPPER UN MODÈLE DE PRÉDICTION ÉCO-RESPONSABLE

RÉSUMÉ

Votre mission est de concevoir un modèle prédictif ayant la plus faible empreinte carbone. Cette compétition est organisée dans le cadre du salon Big Data World Paris les 16 et 17 novembre 2022.

PROJET DATA ÉCO-RESPONSABLE SOBRIÉTÉ NUMÉRIQUE

Citation : Laude, H., & Mamavi, O. (Sep 2022). Sobriété numérique : développer un modèle de prédiction éco-responsable. https://management-datascience.org/challenges/20642/.

Les auteurs :

- Henri Laude (henri.laude@ar-p.com) - Advanced Research Partners
- Olivier Mamavi (omamavi@gmail.com) - Paris School of Business - ORCID : https://orcid.org/0000-0002-6421-1048

Liens d'intérêts : Le ou les auteurs déclarent ne pas avoir connaissance de conflit d'intérêts impliqués par l'écriture de cet article.

Financement : Le ou les auteurs déclarent ne pas avoir bénéficié de financement pour le travail mis en jeu par cet article.

OBJECTIFS

Contexte

Le numérique (ordinateurs, data centers, réseaux…) représente aujourd'hui près de 10% de la consommation d'électricité et 5 % des émissions de gaz à effet de serre au niveau mondial (Diguet et al.,2019). Comme pour les autres secteurs de l'économie, les technologies de l'information vont devoir prendre en compte leur impact environnemental (Trid et al., 2019). La tendance à la surconsommation numérique n'est plus soutenable au

regard de l'approvisionnement énergétique qu'elle requiert. Au-delà des discours, la sobriété numérique devient un impératif (Salas y Mélia, 2022).

La sobriété numérique désigne les changements de comportement qui permettent, notamment, une diminution des consommations d'énergie. La démarche consiste à concevoir des services numériques plus sobres et possédant une faible empreinte carbone. Une entreprise qui s'inscrit dans une démarche de sobriété numérique doit prendre en compte l'impact du stockage des données et des applications dans des serveurs distants. Elle doit également tenir compte des puissances de calcul nécessaire pour le traitement et l'exploitation des données.

Pour apporter une réponse concrète qui permet de réduire le coût énergétique et environnemental d'un projet data, Management & Datascience organise une compétition dans le cadre du salon Big Data World du 16 et 17 novembre à Paris.

Challenge

Vous devrez concevoir un modèle prédictif ayant la plus faible empreinte carbone.

L'objectif est de prédire la classification des attributs de nœuds d'un graphe de connaissance. un graphe est un ensemble d'entités reliées entre elles, composé de nœuds qui représentent les entités et d'arcs ou arêtes qui représentent les relations. Les relations ou arcs peuvent être enrichis par des attributs ou encore une valeur quantitative représentant le poids de la relation. Les graphes de connaissances sont très utilisés par les moteurs de recherche, les services de question-réponse en ligne, les assistants personnels intelligents ou les médias sociaux.

Dans le cadre de ce challenge, il faudra optimiser un modèle capable d'inférer le type d'un article en fonction des articles qui le citent ou qu'il cite. Vous devrez prendre en compte les termes employés respectivement par l'article et les articles liés par le réseau de citation. Il s'agit d'un problème de classification de documents représentés sous la forme d'un graphe.

Le thème précis du challenge sera fourni lors du démarrage du challenge. Vous aurez accès à notre Datalab, dans lequel vous trouverez les données et à une console Jupyter pour réaliser votre modèle. Un programme de référence vous sera également fourni ainsi que des ressources pédagogiques.

Livrable

La structure du programme à livrer est le suivant :

1. lecture des données,
2. *split* aléatoire des données 50/50 en données train et test,
3. création de votre structure de représentation des graphes train et test,
4. démarrage d'un compteur de temps (timer),
5. création et training de votre modèle,
6. calcul de l'accuracy sur les tests et mise en mémoire des prédictions sur l'ensemble du graphe test,
7. fin du comptage de temps,
8. impression de 10 prédictions, comme sur dans le présent code avec les prédiction, mais aussi les probabilité des différentes classes,
9. impression de l'accuracy et du temps passé en dernière ligne et dans le même format que celui produit dans le code de référence présenté plus bas.

Après avoir exécuté leur code dans le datalab, les candidats devront documenté le code pour qu'il soit compréhensible, puis soumettre le script sur la plateforme de Management & Datascience.

Evaluation

La performance de votre modèle sera évaluée en fonction du rapport entre l'efficacité de prédiction et la consommation de puissance informatique lors du traitement du graphe. Les indicateurs de l'évaluation seront les suivants :

1. l'*accuracy* après un *split* aléatoire 50/50 sur les données (calculée par le soumissionnaire)
2. le temps de traitement (calculé par le soumissionnaire)
3. la mémoire maximale utilisée (qui sera mesurée par nos soins)

Nous exécuterons votre code 10 fois et les moyennes de chacune de ces trois mesures seront calculées pour générer le classement technique de votre soumission.

Attention, une note sera également établie sur la **qualité de votre code**. Les critères d'évaluation de la qualité du code seront la clarté et la reproductibilité du script.

Prix

- Les finalistes seront inviter à présenter leurs résultats pendant le salon Big Data World Paris le 17 novembre 2022 à 10h30, lors d'une master class. Des experts scientifiques et professionnelles discuterons les résultats pour comprendre comment développer un projet data éco-responsable.
- Les vainqueurs recevront le trophée du **meilleur datascientist 2022 pour l'environnement**.
- Les vainqueurs bénéficieront d'une couverture médiatique (presse et réseaux sociaux) ainsi que la publication de leurs résultats au sein de la revue scientifique **Management & Data Science**

Références

- Diguet, C., Lopez, F., & Lefèvre, L. (2019). *L'impact spatial et énergétique des data centers sur les territoires* (Thèse de doctorat, ADEME, Direction Villes et territoires durables).
- Salas y Mélia, D. (2022). Les principaux enseignements du 6e rapport du groupe I du GIEC. *Annales des Mines – Responsabilité et environnement*, 106, 11-16. https://doi.org/10.3917/re1.106.0011
- Trid, S., Corbett, J., & Bouchard, L. (2019). Modèle théorique de projets de Green IS: une spécification des relations entre objectifs, compétences et culture environnementale. *Systemes d'information management*, *24*(1), 7-45.

MODALITÉS

Participants

Ce data challenge est ouvert à tous, notamment :

- aux étudiants en cycle master d'écoles d'ingénieur, de commerce ou des universités
- aux chercheurs
- aux datascientists

Inscription

La participation au data challenge est **gratuite**.

Chaque candidat doit s'inscrire au préalable sur le site web de **Management & Data Science** , puis constituer une équipe avec des membres de la communauté déjà inscrits.

Le nombre de membres maximum par équipe est de trois (3) personnes.

Datalab

Management & Datascience fourni à chaque équipe un espace de travail à utiliser obligatoirement au sein de son laboratoire de données.

Chaque équipe aura une console Jupyter avec python et R, Tensorflow et Pytorch plus conda et pip3, les tidyverse r, caret r, pycaret, statsmodels, base r, mlr r, fs r, r datatable, devtools pour r, reticulate r, dplyr r, stringr r, readr r, scikit-learn, numpy, scipy, pandas avec toutes leurs dépendances.

Jury

L'ensemble des propositions sera évalué et classé en fonction entre du rapport **accuracy/sobriété et de la qualité du code fourni**.

Les 2 meilleures propositions iront en finale. Les candidats finalistes présenteront leurs résultats devant un jury lors du salon Big Data World Paris le jeudi 17 novembre 2022 à 10h30 pendant 7 minutes.

Le jury désignera le vainqueur en classant les meilleures propositions.

Le jury est composé des membres suivants :

- **Alexandre ALFOCEA**, Consultant Data Science (LiveRamp)
- **Henri LAUDE**, Chief Data Scientist (Advanced Research Partners)
- **Olivier MAMAVI**, Professeur en Data Management (Paris School of Business)
- **Joël TANKEU**, Associate Solutions Architect (Amazon Web Services)

DONNÉES

Les données à utiliser pour ce challenge, ainsi qu'un programme de référence (fourni par Henri Laude) et la documentation nécessaire pour concevoir le modèle prédictif seront disponible lors du démarrage du challenge.

RÈGLEMENT

Conditions générales

La participation au challenge implique pour tout participant **l'acceptation entière et sans réserve des règles ci-dessous**. Le non-respect dudit règlement entraîne l'annulation immédiate de la participation.

1. L'inscription et la participation au challenge est entièrement gratuite et libre.
2. L'inscription au challenge doit se faire de manière individuelle. Afin de participer au challenge le participant doit avoir créé un compte utilisateur sur le site de Management & Data Science, et renseigné de manière loyale et complète les informations requises, telles que le nom, prénom, adresse mail, etc.
3. Les participants individuels peuvent choisir de former une équipe de deux (2) à trois (3) membres maximum pour soumettre leur livrable.
4. Le fait, pour un participant ou une équipe, de ne pas déposer avant la date limite le livrable sur le site du challenge sera considéré comme un abandon de sa/leur part au challenge. Le participant ou l'équipe ne pourra en aucun cas réintégrer le challenge.
5. Les contributions sont publiées sous une licence Creative Commons **Attribution/Pas de modification.**
6. Chaque soumission sera notée et classée selon la métrique d'évaluation indiquée sur le site Web du concours. Le(s) gagnant(s) potentiel(s) seront avisés par courriel.
7. Les données personnelles du participant font l'objet d'un traitement au sens de la réglementation sur la protection des données personnelles (Règlement (UE) 2016/679 du parlement européen et du conseil du 27 avril 2016 dit « RGPD ») pour lequel Management & Data Science définit les finalités et les moyens et est, à ce titre, responsable de ce traitement au sens du RGPD.
8. Management & Data Science ne saurait être tenue responsable de toutes perturbations, à la fois sur le réseau internet ou des difficultés d'accès liées à un grand nombre de connectés ou de participants. Management & Data Science ne peut en aucune manière être tenue responsable des coupures de communication ou d'accès, des pertes de données, des virus informatiques ou de tout préjudice direct ou indirect quel qu'il soit, éventuellement subi par un participant avant pendant et après sa participation au challenge. En conséquence, les participants renoncent à tout recours contre Management & Data Science et ses préposés pour des dommages et/ou préjudices qu'ils pourraient subir dans le cadre du challenge.