Challenge on Data Quality

Comment évaluer de manière automatique la qualité d'une table de données ?

Définition du sujet :

Les données sont la pierre angulaire de Foyer et sont à la base de nombreuses décisions. Cependant, pour prendre de bonnes décisions, il faut des données de qualité. On estime qu'une donnée est de qualité si elle est précise, exhaustive, fiable et d'actualité. Actuellement, le seul moyen que nous ayons pour améliorer la qualité des données passe par des règles deterministes. Le but de ce challenge serait d'avoir des indicateurs sur la qualité des données adaptables à tout type de tables sans connaissances métier et sans fixer des règles a priori ni réaliser ces contrôles a posteriori. On peut penser par exemple à un indicateur pour savoir combien il y a d'erreurs sur une colonne d'une table de données.

Données d'entrée

Il vous sera proposé de tester vos algorithmes sur une table de données de 100000 lignes au format csv fournie par une entreprise de vente de machines agricoles. Cette table contient le détail des ventes des machines ainsi que des données précisant les caractéristiques de ces dernières. De plus, il vous sera fourni un dictionnaire de données donnant une explication de chaque colonne de la table.

Livrables

- Proposition d'une definition de la qualité d'une donnée
- Algorithme (descriptif + code source) qui permet d'évaluer la qualité d'une donnée et de l'ensemble des données d'une table
- Bilan des erreurs de la table fournie en entrée comportant pour chaque erreur son type ainsi que son ID

Bibliographie

https://www.win.tue.nl/~mpechen/publications/pubs/Gama_ACMCS_AdaptationCD_accepted.pdf

https://mobidev.biz/blog/unsupervised-machine-learning-improve-data-quality

https://towardsdatascience.com/automated-data-quality-testing-at-scale-using-apache-spark-93bb1e2c5cd0

https://www.vldb.org/pvldb/vol11/p1781-schelter.pdf

https://arxiv.org/ftp/arxiv/papers/1810/1810.07132.pdf

https://res.mdpi.com/d_attachment/symmetry/symmetry-10-00248/article_deploy/symmetry-10-00248.pdf

https://arxiv.org/ftp/arxiv/papers/2009/2009.06672.pdf