|  |  |  |
| --- | --- | --- |
| Social Nantes | 10 mars  2017 | |
| Ce document concerne l’apprentissage du Machine Learning autrement dit « Data Mind » ou « Intelligence Artificielle » au gré des visions et des époques | |  |

Sommaire

[1) Présentation du Machine Learning\* 3](#_Toc477087328)

[2) Introduction au projet Social-Nantes 3](#_Toc477087329)

[3) Démarche / Méthode 3](#_Toc477087330)

[4) Mise en œuvre 3](#_Toc477087331)

[5) Difficultés rencontrées 3](#_Toc477087332)

[6) Résultats obtenus 4](#_Toc477087333)

[7) Solutions apportées 4](#_Toc477087334)

[8) Leçons à en tirer 4](#_Toc477087335)

[9) Pour aller plus loin 4](#_Toc477087336)

[10) Conclusion 4](#_Toc477087337)

# Présentation du Machine Learning\*

Le machine learning consiste, avec des bases d’algorithmes mathématiques complexes, à apprendre à une machine à apprendre. Aussi nommée « Data Mining » ou « Intelligence artificielle», en fonction des lieux et des époques. Cette discipline a rapidement évolué ces 10 dernières années en raison de l’augmentation de la puissance de calculs de nos ordinateurs et de la multiplication des données disponibles. A quoi cela sert-il ? Que peut-on en faire ? Comment l’utiliser ? Ce qu’il faut faire ou ne pas faire, les difficultés que l’on peut rencontrer, comment les contourner et les questions à se poser sur le sujet seront abordés dans ce rapport.

# Introduction au projet Social-Nantes

La loi numérique promulguée le 7 octobre 2016 à permit l’ouverture des données publique donnant accès au public à une large mise à disposition de ces données. Nantes est un acteur pionnier puisque l’Open Data de la ville est particulièrement fourni avec ses 834 bases de données rendues publiques. Travaillant à Nantes Métropole Habitat je me suis penchée sur les données sociales de la ville. L’idée était de voir puis de prédire s’il y a un lien par commune entre le nombre de bénéficiaire du RSA (Revenu Social d’Activité) et le nombre de bénéficiaire du FSL (Fond de solidarité logement). En prédisant que plus une commune a de bénéficiaires du RSA, plus elle a de bénéficiaires du FSL.

# Démarche / Méthode

Afin de réussir à mener ce projet à bien il me fallait donc :

* Récupérer les données sur le site : <http://data.nantes.fr/donnees/>
* Importer les bibliothèques nécessaires à Python
* Lire les données
* Les nettoyer
* Les mettre en forme
* Les analyser
* Les traiter et effectuer des calculs mathématiques (regression linéaires et coefficient de corrélation).
* En tirer les conclusions

# Mise en œuvre

* L’import des bibliothèques ne pose pas de problème
* Trouver les données et les importer dans le projet non plus.
* Le nettoyage s’avère plus compliqué à mettre en œuvre
* Sans nettoyage impossible d’analyser les données à traiter
* Sans nettoyage pas de possibilité d’effectuer des calculs sur des cellules vides
* Sans calculs impossible d’en tirer quelque conclusion que ce soit

Le projet est visible ici : <https://github.com/AnnetteHellebois/2016-2017/blob/master/ProjetFinal/Projet_Final_V1.ipynb> on peut constater que le nombre de bénéficiaires du FSL est peu renseigné et que le nettoyage des données n’a pas été fait correctement ce qui impacte la suite du projet.

# Difficultés rencontrées

* L’outil : Git s’avère ardu à prendre en main, tout se passe en ligne de commandes, outil que je maîtrise peu puisque ce sont les fenêtres qui m’ont menée à l’informatique
* La programmation en langage Python, lorsqu’on n’en n’a jamais n’est pas simple non plus
* Les cours sont donnés en français mais les supports de cours eux sont en anglais, ce qui ajoute une difficulté supplémentaires
* Les formules mathématiques sont difficilement lisibles quand on a peu fait de mathématiques, impossible de savoir ce que fait telle ou telle formule.

# Résultats obtenus

Les résultats obtenus sont peu probants. Si la recherche des données s’est révélée simple, leur qualité laisse à désirer. Le même problème se pose à mon travail à Nantes Métropole Habitat qui compte 700 personnes mais 300 métiers différends. Renseigner TOUTES les données est primordial à leur analyse. Des techniques existent pour pallier à ce problème comme remplacer les données manquantes par la médiane par exemple. Mais avec peu de notions de programmation c’est un peu plus complexe que cela en a l’air.

# Solutions apportées

Concernant Git : GitKraken s’avère plus graphique et plus intuitif que git ou Tortoise, seul le paramétrage est difficile à réaliser mais une fois que c’est fait il est très intuitif.

Concernant la programmation en Python, c’est en forgeant qu’on devient forgeron. 100 sur le métier j’ai remis mon ouvrage, à en attraper des insomnies. ..

Concernant les mathématiques : Les cours de statistiques suivi après le module de Machine learning m’ont permit de mieux appréhender les formules et notions mathématiques.

Concernant l’anglais : en suivant des tutoriels j’ai aussi pu comprendre les notions complexes.

De nombreux tutoriels sont disponibles ici : http://eric.univ-lyon2.fr/~ricco/cours/supports\_data\_mining.html

# Leçons à en tirer

Il ne faut jamais désespérer, il faut être courageux, persévérer, communiquer et des progrès sont toujours possibles tant qu’on n’abandonne pas.

# Pour aller plus loin

Le projet est perfectible, loin s’en faut. Plutôt que de mesurer le nombre de bénéficiaire de tel ou tel fond j’aurais pu comparer le taux, à savoir le nombre de bénéficiaire par rapport au nombre total d’habitants de la commune ce qui aurait permit une vue plus réaliste. Hélas le site de l’Insee était indisponible lorsque j’ai souhaité faire cette étude. Les données sont maintenant localisées, elles se trouvent ici : <https://www.insee.fr/fr/statistiques/2409376?sommaire=2409559>

D’autre part pour un rendu plus visuel j’aurais pu faire des graphiques à l’aide de la librairie MathPlotLib.

# Conclusion

Se familiariser avec de nouveaux outils n’est jamais aisé, faire de nouveaux apprentissages chaque jour, combler ses lacunes, tenter de rattraper son retard peu parfois se montrer décourageant. Mais qu’il est enrichissant d’apprendre toujours et encore.

Toutefois n’oublions pas l’éthique dans tout cela. Qui est responsable en cas de problème lorsque par exemple une voiture qui se conduit toute seule est impliquée dans un accident ? Que se passerait-il si l’une d’elles finissait par se montrer agressive envers l’homme lorsqu’elle se sent menacée ? En effet une étude sur le comportement des IA a montrée qu’elles pouvaient devenir agressives pour obtenir ce qu’elles voulaient. Un article de l’express en conclut : « que les chercheurs devront être prudents lors de la conception des robots de demain et s’assurer que les IA sont récompensées quand elles font passer les intérêts humains avant les leurs. » [lire l’article](https://fr.express.live/2017/02/22/intelligence-artificielle-ia-google-fortement-agressive/)

Il est aussi temps de se poser les questions de cet ordre.