American Journal of Sciences and Engineering Research

E-ISSN -2348 – 703X, Volume 5, Issue 6, 2022

Extraction of Knowledge from Civil Status Data (Surname and First Name) By Artificial Intelligence

Koto J.B[1]– Ramahefy T.R[2]– Randrianja S.[3]

École Doctorale Glocalisme, Environnement et Sécurité des Sociétés indienocéaniques (GENESIS) - Université d'AntsirananaAntsiranana 00201 – Madagascar

Resume :Cet article propose un modèle d'extraction de connaissance à partir des données de l'état civil tel que les « noms et prénoms ». Notre objectif est de démontrer que la fouille de textes doit nécessairement exploiter un modèle de connaissance. Nous avons pris comme une exemple une liste des « nom et prénom » comme le domaine à étudier. Par l'intelligence artificielle et les outils d'analyse de python, des résultats peuvent se présenter sous forme de graphique. Les résultats montrent que la formulation des « nom et prénom » à Madagascar sont faites librement. Plusieurs échantillons sont semblables alors d'autres sont très différents.

MOTS-CLÉS : Intelligence artificielle - Fouille de textes - extraction des connaissances, détection de doublon

I. Introduction

La fouille des données est un sujet très connu aujourd'hui dans le monde de la science des données. Des nombreuses publications parlent de la fouille de textes. En effet, c'est la science qu'on peut utiliser dans plusieurs études pour l'extraction de la connaissance à partir des données. Nous présentons dans cet article une méthode d'extraction de connaissance à partir des données de l'état civil, les « nom et prénom » par l'intelligence artificielle.

En outre, Python est un langage de programmation libre de droits et utilisé dans plusieurs domaines. Il est utilisé pour développer des applications, des interfaces graphiques d'outils, ou encore pour faire du développement logiciel en général, et aussi dans le domaine du *Data Science*.

II. Définition et processus de la fouille de textes

2.1. Définition

La fouille de textes, traduit du terme en anglais *text mining,* est apparue dans la deuxième moitié des années 90, en écho à des travaux réalisés depuis les années 80 sur des bases de données. En 1991, Piatesky-Shapiro introduit comme titre de son ouvrage [1] le terme de *Knowledge Discovery from Databases,* abrégé par la suite en KDD et, dont l'équivalent français est *Extraction de Connaissances à partir de Bases de Données* (ECBD). Ce n'est que vers 1995 que l'usage des termes *Knowledge Discovery from Databases* et *Data Mining* se précise. L'extraction de connaissances à partir de la base de données désigne alors le processus global de découverte de connaissance qui permet de passer de données brutes à des connaissances alors que la fouille de données n'est qu'une étape de l'ECBD au cours de laquelle un modèle est construit.

Extraction de connaissance à partir de textes

 Received-20-11-2022, Accepted- 04-12-2022

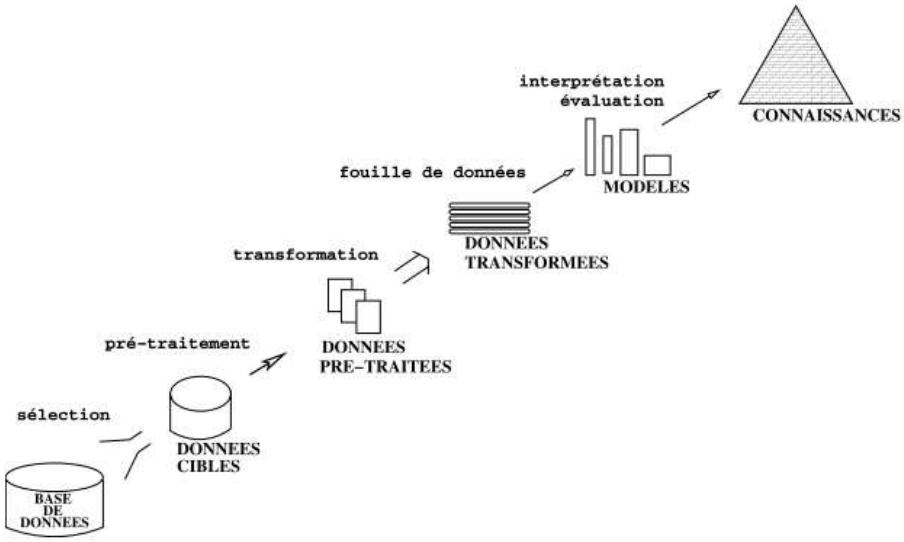


Schéma global de l'ECBD [2b]

L'ECBD peut se décomposer en de nombreuses étapes plus ou moins complexes mais la figure précédente en donne une vision synthétique. Parmi les grandes étapes de l'ECBD, on peut distinguer :

- la **sélection** qui crée un ensemble de données à étudier ;
- le **prétraitement** qui vise à enlever le bruit et à définir une stratégie pour traiter les données manquantes ;
- la **transformation** où l'on recherche les meilleures structures pour représenter les données en fonction de la tâche ;
- la **fouille de données** proprement dite et la définition de la tâche basées sur la classification, la recherche de modèles… et la définition des paramètres appropriés ;
- l'**interprétation et l'évaluation** pendant laquelle les patrons extraits sont analysés. La connaissance qui en est ainsi extraite est alors stockée dans la base de connaissances.

La définition de l'ECBD montre qu'elle entretient des liens forts avec l'apprentissage.

Nous calquons donc notre définition de l'extraction de connaissances à partir de textes sur celle de l'extraction de connaissances à partir de bases de données.

2.2. Processus de fouille

Les processus mis en œuvre pour la fouille dans les textes exploitent très largement les méthodes et techniques mises en œuvre plus globalement sur des données. Nous nous intéressons ici à la connaissance générale qui sont cachés dans les textes, « noms et prénoms ». La fouille de textes peut se faire qu'en s'appuyant sur des connaissances du domaine.

Ainsi, l'enrichissement progressif du modèle de connaissances par les itérations successives de la boucle d'extraction de connaissance à partir du texte (ECT) permet d'extraire de nouvelles connaissances et d'enrichir de nouveau le modèle. Les règles d'association ainsi que certaines méthodes de classification symbolique comme les treillis de Galois en analyse formelle des concepts peuvent être plus facilement associées à un modèle de connaissance que certaines autres méthodes numériques [3].

De par leur lisibilité, les règles d'association constituent une méthode de fouille attractive. Nous donnons brièvement une définition des règles d'association puis mentionnons quelques travaux sur leur utilisation en fouille de textes.

Nous introduisons les règles d'association dans le contexte de la fouille de textes. Chaque texte est représenté par un ensemble de concepts qui modélise son contenu [3].

Soit $T = \{t_1,\ldots..,t_n\}$ l'ensemble fini non vide de textes. Soit $C = \{c_1,\ldots..,c_n\}$ l'ensemble des concepts modélisant les textes *T*.*T* et *C* sont lié par une relation $R \subseteq T \times C$.

Une règle d'association est une implication de la forme $B \rightarrow H$ où B est la prémisse (body), et H est la conclusion (head) avec $B \subseteq C$, $H \subseteq C$ et $B \cap H = \emptyset$.

Si $B = \{c_1,....c_p\}$ est l'ensemble des concepts de la prémisse d'une règle d'association r_j *et* $H = \{c_{p+1},....cq\}$ l'ensemble des concepts de la conclusion de r_j, $B \rightarrow H$ signifie que tous les textes de T contenant les concepts c_1, c_2,.....c_p contiennent aussi les concepts c_{p+1}, c_{p+2},....,c_q avec une certaine probabilité P.

Le support de r_j est le nombre de textes contenant les concepts de B. La confiance de r_j est le rapport entre le nombre de textes contenant les concepts B U H ($\{ci_1,...,ci_p,...ci_q\}$) et le nombre de textes contenant H ($\{ci_1,...,ci_p,...ci_q\}$). Ce rapport est défini par la probabilité conditionnelle P (B|H). Le support et la confiance sont deux mesures associées aux règles d'association [5] et exploitées par les algorithmes d'extraction de règles pour en réduire la complexité. Deux valeurs de seuils sont alors définies σ_s pour le support minimal et σ_c pour la confiance minimale [3].

III. Présentation du corpus

Nous avons réalisé pour cette expérimentation un corpus composé de 1000 enregistrements des individues (noms et prénoms), de nationalité malagasy, résident à Madagascar.

Le prénom et le nom de famille sont les premiers éléments que nous utilisons pour identifier et se faire identifier au sein de la société. Le nom de famille nous rattache à une certaine lignée (par la filiation) tandis que le prénom laissé au libre choix des parents permet de s'individualiser. Les parents sont libres aussi de donner le nombre des prénoms qu'ils souhaitent. D'autres parents donnent un ou deux ou plus des prénoms, tandis que certain les préfèrent de lui donner qu'un nom, aucun prénom. Ce qui fait qu'un individu a toujours un nom et sans prénom, ou avoir un ou plusieurs prénoms.

IV. Application et présentation des résultats de l'extraction d'information

L'outils d'analyse utilisé pour notre application est basé sur la librairie standard de python. Il fournit des API orientées objet pour l'intégration de tracés dans les applications. Il est similaire à MATLAB en termes de capacité et de syntaxe.

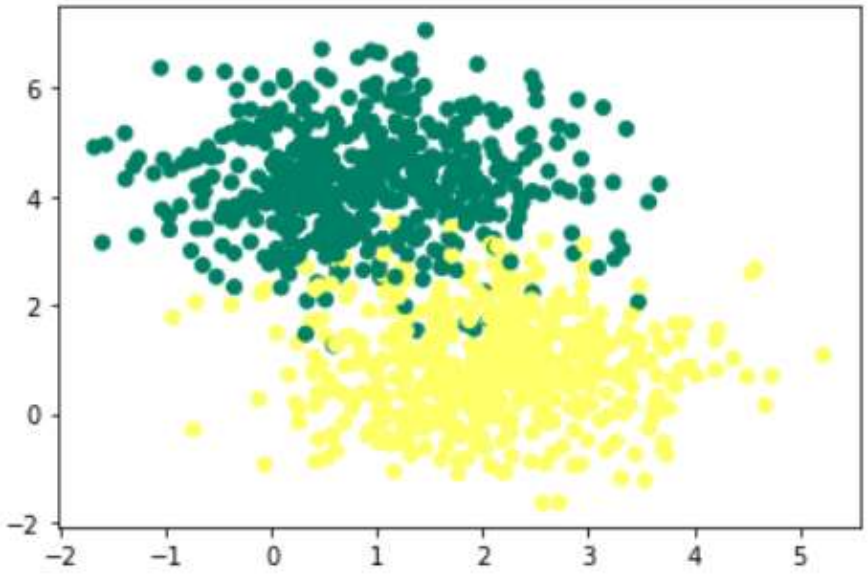
La première tâche est d'importer les bibliothèques *numpy* et *matplotlib*.

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.datasets import make_blobs
```

4.1. Dataset

Nous pouvons aperçus dans la figure ci-après, nous avons 1000 variables dans notre dataset.et nous montre bien que c'est 1000 réponses que le neurone nous donne.

```
dimensions de X: (1000, 2)
dimensions de y: (1000, 1)
```



4.2. Fonctions du modèle

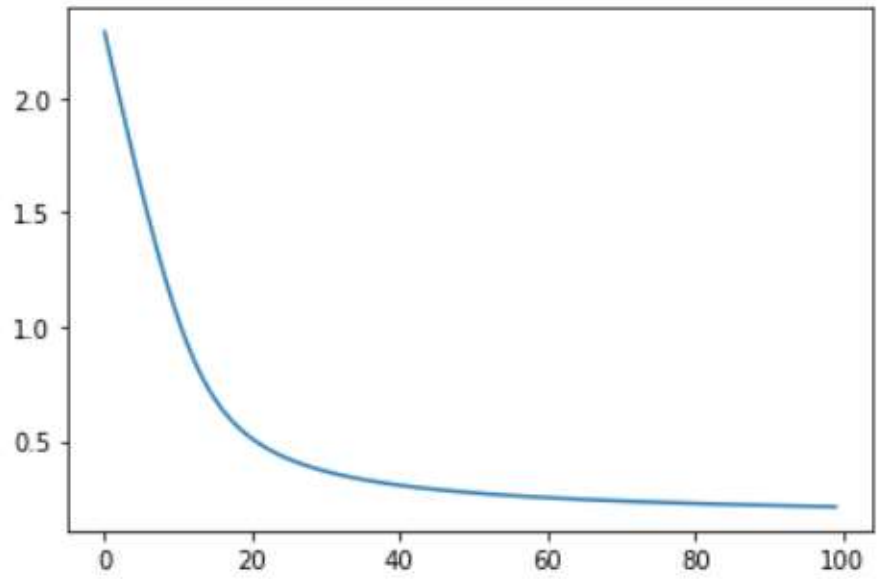
En construisant d'une fonction pour le modèle d'apprentissage, nous obtient la figure ci-après :

```
                                    def predict(X, W, b):
    A = model(X, W, b)

    return A >= 0.5
from sklearn.metrics import accuracy_score
def artificial_neuron(X, Y, learning_rate = 0.1, n_iter = 100):

    W, b = initialisation(X)
    Loss = []
    for i in range(n_iter):
        A = model(X, W, b)
        Loss.append(log_loss(A, Y))
        dW, db = gradients(A, X, Y)
        W, b = update(dW, db, W, b, learning_rate)
    y_pred = predict(X, W, b)
    print(accuracy_score(y, y_pred))
    plt.plot(Loss)
    plt.show()
    return (W, b)
W, b = artificial_neuron(X, Y)
```

```
0.934
```

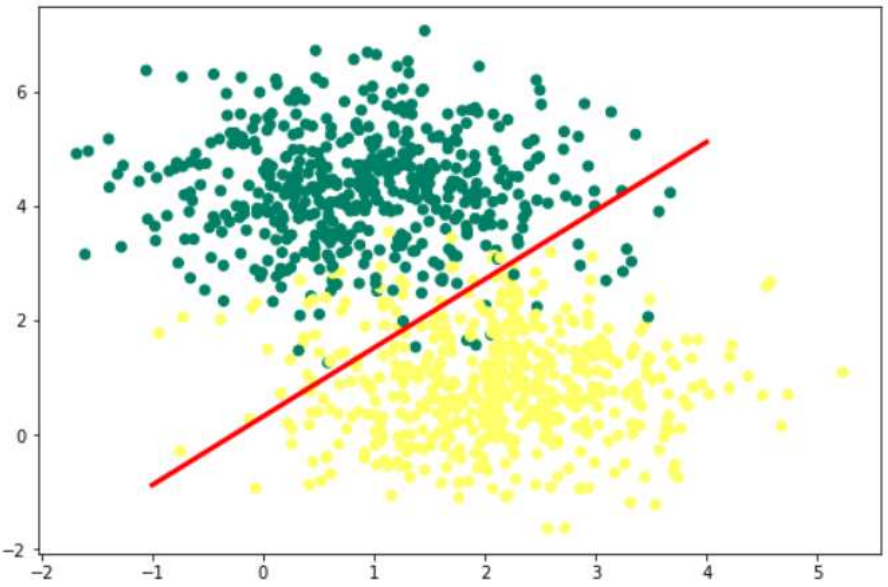


En apercevant notre courbe d'apprentissage ci-dessus, l'évolution des erreurs effectués par le model au fur et à mesure que celle-ci apprends. Les erreurs diminuent et la fonction coût converge vers une valeur plateau. c'est la valeur le plus bas.

4.3. Frontière de décision

En comparant les variables et la prédiction, nous mettrons la frontière de décision. La frontière de décision est l'endroit pour lesquels la probabilité est supérieur à 50%. C'est l'ensemble des points ou Z= 0 c'est-à-dire a(Z) = 0,5.

```
fig, ax = plt.subplots(figsize=(9, 6))
ax.scatter(X[:,0], X[:, 1], c=y, cmap='summer')
x1 = np.linspace(-1, 4, 100)
x2 = ( - W[0] * x1 - b) / W[1]
ax.plot(x1, x2, c='red', lw=3)
```



Les points variables au-dessus de la ligne en rouge, ce sont des variables qui présentes des informations les plus proches entre eux alors que les variables au-dessous de la ligne, ce sont des variables qui sont très différents entre eux. C'est-à-dire ce sont des « nom et prénom » qui sont conçus différemment.

V. Conclusion et perspective

L'intérêt d'analyse décrite est d'apporter des connaissances sur la formulation des « nom et prénom » à Madagascar. Les résultats sont satisfaisants puisque nous avons réussi à corréler les données entre eux. D'autres sont très proches entre eux tant dis qu'il existe des « nom et prénom » qui sont très différents par rapport aux autres noms. Cela nous prouve que les noms à Madagascar sont conçus d'une manière libre. L'intelligence artificielle est une science qui s'applique dans tous les domaines. La perspective de programmer une intelligence artificielle s'augment de plus en plus après avoir obtenir des résultats du recherche et découvrir son importance dans la science.

VI. Références

[1] Piatetsky-Shapiro G., Frawley W., *Knowledge Discovery in Databases*, AAAI Press, 1991.

[2] Fayyad U. M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R., Eds., *Advances in Knowledge Discovery and Data Mining*, AAAI Press, 1996.

[2] Fayyad U., Piatetsky-Shapiro G., Smyth P., *« Froma Data Mining to Knowledge Discovery »*, chapitre1, Fayyad et al. [2a], 1996.

[3] Yannick Toussaint, *Extraction De Connaissances À Partir De Textes Structurés*, Lavoisier « Document numérique », 2004/3 Vol. 8 | pages 11 à 34.

[4] Amandine VELT, *Python pour la Data Science*, 381 pages.

[5] Andre J., Futura R., Quint V., Eds., Structured documents, The Cambridge Series on electronic publishing 1989.