American Journal of Sciences and Engineering Research

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

Artificial Intelligence Applied to the Process of Analysis of Recruitment Results

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

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

---

Résumé :

Nous présentons l'application d'analyse des résultats d'entretient par l'intelligence artificielle. L'accent est mis sur le langage de programmation Python. Celle-ci dispose des librairies tels que NumPy, Pandas et Matplotlib qui sont utiles à l'analyse de données. L'analyse de donnée ou la science de donné est un moyen, pour les entreprises,

 Received-05-10-2022, Accepted- 19-10-2022

d'avoir un aperçu du fait sur les activités et aussi de planifier les projets. C'est un travail scientifique que le dirigent le confier à l'analyste.

Mots clés : Intelligence artificielle, analyse de données, entretien d'embauche

---

I. Introduction

L'étude de données est une science qui dévient plus en plus pratiquer par des diverses entreprises que se soient dans le domaine de production ou dans le domaine de marketing et de commerce des produits. L'objectif est toujours dans le perceptive d'amélioration du système encours pour un but déterminé ou pour l'étude d'un nouveau projet. Cet article traite le résultat d'entretien d'embauche, qui est notre donnée ici, avec l'application de l'intelligence artificiel. Notre objectif est de pouvoir analysé le donné d'une manière scientifique. Les résultats seront présentés sous forme de tableau et des figures avec quelles phrases d'interprétation des illustrations.

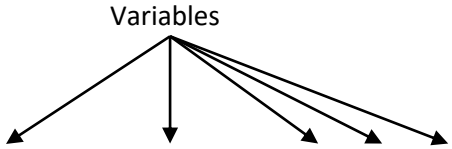
En outre, Python est un langage de programmation libre de droits et utilisé dans plusieurs de domaines. Il est utilisé pour développer des applications web, des jeux, des interfaces graphiques d'outils, ou encore pour faire du développement logiciel en général, et aussi dans le domaine scientifique pour l'analyse de données, et plus généralement la *Data Science*. Les statistiques montrent que plus de 66% des scientifiques des données utilisent quotidiennement Python et 84% l'utilisent comme langage principal [1].

II. Analyse exploratoire (descriptive)

Une variable est une propriété ou caractéristique d'un individu.

- Exemple : Couleur des yeux d'une personne, âge, état civil, ...
- Une collection de variables décrivant à un individu

On dit individu ou enregistrement, point, cas, objet, entité, exemple l'extrait de données utilisées dans cet article.



sexe	parcours	age	niv	note
F	HISTOIRE	27	L3	8
G	TOURISME	19	L2	5
F	ENVIRONNEMENT	19	L2	6
F	GEOGRAPHIE	19	L2	7
F	MATHS	24	L3	8

Types de variables :

- Qualitative : les variables représentent des catégories différentes au lieu des numéros. Les opérations mathématiques comme la somme et la soustraction n'ont pas de sens. Exemples : sexe, parcours, niv (niveau académique)
- Quantitative : les variables sont les numéros. Exemple : age, note

Transformation d'une variable quantitative en variable qualitative

Pour les variables discrètes : considérer que les valeurs prises par la variable sont les modalités de la variable qualitative (ordonnée).

Pour les variables continues [4]:

- on divise l'intervalle [*a* ; *b*[où varie la variable en un certain nombre d'intervalles [*a* ; x_1[, [x_1 ; x_2[, [x_i; x_{i+1}[… , [x_{p-1}; *b*[et
- on dénombre pour chaque intervalle le nombre d'individus dont la mesure appartient à l'intervalle.
- En règle générale, on choisit des classes de même amplitude.
- Pour que la distribution en fréquence soit intéressante, il faut que chaque classe comprenne un nombre « suffisant » d'individus (*ni*).
- Si la longueur des intervalles est trop grande, on perd trop d'information.

Il existe des formules empiriques pour établir le nombre de classes pour un échantillon de taille *n*

- Règle de Sturge : Nombre de classes =1 + 3.3 log *n*
- Règle de Yule : Nombre de classes = 2.5√n
- L'intervalle entre chaque classe est calculé par (*b*-*a*)/nombre de classes
- On calcule ensuite à partir de *a* les classes successives par addition.

NB: Il n'est pas obligatoire d'avoir des classes de même amplitude, mais pas de chevauchement d'intervalle.

Les données

Le point de départ est d'une table de données :

$$X = \begin{pmatrix} x_{11} & x_{1j} & x_{1m} \\ x_{i1} & x_{ij} & x_{im} \\ x_{n1} & x_{nj} & x_{nm} \end{pmatrix}$$

individue i

individue j

Description d'une variable quantitative

Une variable quantitative est décrite par les valeurs qui prennent l'ensemble de *n* individus pour lesquels a été définis [4].

Pour résumer l'information d'une variable quantitative les indices les plus communes sont [4]:

- La moyenne est définie :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- La variance est définie :

$$var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

- L'écart type :

$$\sigma_x = \sqrt{var(X)}$$

- Le Coefficient de détermination :

R^2 = Var (estimés par l'équation de régression) / Var (totale)

$$R^2 = \frac{var(aX + b)}{var(Y)}$$

- Le Coefficient de corrélation :

$$R = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

Matrice de Corrélation

La grande corrélation positive implique que si une variable augmente l'autre aussi augmente. La grande corrélation négative implique que si une variable augmente l'autre diminue et vice versa. La corrélation proche de 0 implique l'absence de relation entre les variables [4].

III. Outil d'analyse et structure des données

NumPy est la librairie Python dédiée au calcul scientifique et des structures de données [3]. En Data Science, il est essentiel d'avoir des structures adaptées pour stocker et manipuler de grandes quantités de données. C'est là qu'intervient *NumPy*, qui intègre une nouvelle structure de données en Python, les *ndarrays* (tableaux à N dimensions), qui sont des tableaux multidimensionnels ou matrices.

Pandas est un package Python fournissant des structures de données pour rendre le travail avec des données « relationnelles » ou « étiquetées » à la fois simples et intuitives. En voici un bref aperçu de ses importances [2] :

Objets : les classes Series et DataFrame ou table de données.

Lire, écrire création et exportation de tables de données à partir de fichiers textes (séparateurs, .csv, format fixe, compressés), binaires (HDF5 avec Pytable), HTML, XML, JSON, MongoDB, SQL...

Gestion d'une table : sélection des lignes, colonnes, transformations, réorganisation par niveau d'un facteur, discrétisation de variables quantitatives, exclusion ou imputation élémentaire de données manquantes, permutation et échantillonnage aléatoire, variables indicatrices, chaînes de caractères...

Statistiques élémentaires uni et bivariées, tri à plat (nombre de modalités, de valeurs nulles, de valeurs manquantes...), graphiques associés, statistiques par groupe, détection élémentaire de valeurs atypiques...

Manipulation de tables : concaténations, fusions, jointures, tri, gestion des types et formats...

Pandas utilise plusieurs méthodes pour créer des graphiques des données dans le bloc de données. Pour cet article nous choisissons *matplotlib* pour les présentations des graphiques.

IV. Application de l'analyse de données

La première tâche est de l'importer les bibliothèques *numpy*, *pandas* et *matplotlib*. En utilisant le code *pd.read_execel*, nous obtiendrons notre *DataFrame.*

Out[2]:

	num	sexe	parcours	age	niv	note
0	1	F	HISTOIRE	27	L3	8
1	2	G	TOURISME	19	L2	5
2	3	F	ENVIRONNEMENT	19	L2	6
3	4	F	GEOGRAPHIE	19	L2	7
4	5	F	MATHS	24	L3	8
...	...	...	...	...	...	...
514	515	G	INFORMATIQUE	24	L3	3
515	516	F	MATHS	32	M1	9
516	517	G	INFORMATIQUE	33	M1	7
517	518	F	GEOGRAPHIE	33	M1	8
518	519	F	HISTOIRE	34	M1	10

519 rows × 6 columns

Nous avons ici 519 candidats qui ont passé l'entretient : 519 lignes et 6 colonnes.

Voici les points importants de cette analyse :

- **sexe**, F si c'est une fille, G si c'un garçon,
- **parcours** : Il existe des différentes parcours comme nous pouvons voire sur Out[2] : HISTOIRE, TOURISME, ENVIRONNEMENT, GEOGRAPHIE, MATHS, INFORMATIQUE, ….
- **age**, **niv** respectivement (niveau des étudiants dans leur parcours)
- **note** qui est la note obtenue de l'entretient.

Nous commençons notre première analyse par l'utilisation de *describe.*

Out[4]:

	num	age	note
count	519.000000	519.000000	519.000000
mean	260.000000	23.917148	4.697495
std	149.966663	3.399250	1.992024
min	1.000000	17.000000	2.000000
25%	130.500000	22.000000	3.000000
50%	260.000000	25.000000	4.000000
75%	389.500000	26.000000	6.000000
max	519.000000	34.000000	10.000000

describe donne le statistique rapide pour des colonnes de valeurs numériques.

Encore une fois, notre *DataFrame* a 519 lignes de données. Il est vérifié par la première ligne *count*. Nous avons l'âge moyenne de 23,91 ans et la note moyenne de 4,697. Nous apercevons aussi l'âge minimal et maximal qui sont 17ans et 34ans ; et la note minimal et maximal :2 et 10. D'où, la note des candidats varie de 2 à 10.

Un aperçu général pour la colonne sexe, parcours et niveau :

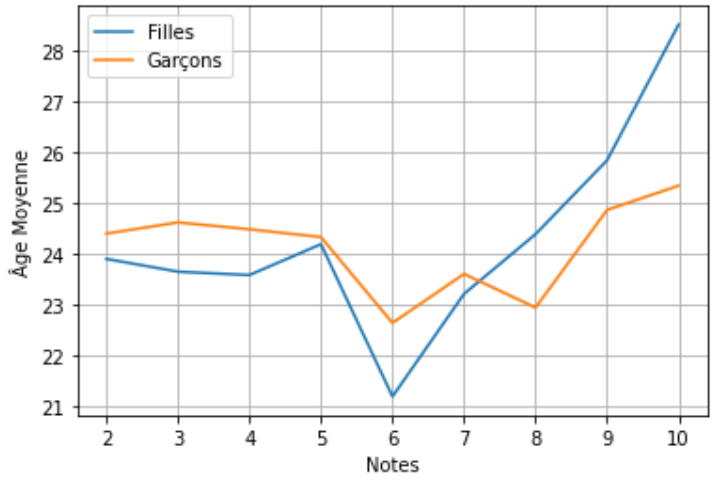
```
Out[5]: F    317
        G    202
        Name: sexe, dtype: int64
```

Les 519 candidats se divisent 317 du sexe féminin et de 202 du sexe masculin.

```
Out[10]: MATHS            68
         GEOGRAPHIE       56
         ECONOMIE         56
         GESTION          55
         INFORMATIQUE     53
         HISTOIRE         49
         ENVIRONNEMENT    48
         DROIT            46
         MEDECINE         46
         TOURISME         42
         Name: parcours, dtype
```

```
Out[12]: L3    318
         L2    177
         L1     19
         M1      5
         Name: niv,
```

Pour la note, du sexe et l'âge moyenne des candidats :



Pour le sexe fille, la note varie de 2 à 10. La moyenne d'âge de la note minimale est de 21,19 ans et de 28,5ans pour la note maximale.

On remarque que les candidates qui ont eu la note 6 ont l'âge moyenne minimale, 23 ans. Pareille pour le sexe garçon, les candidats qui ont la note 6, ont l'âge moyenne minimale de 22,63 ans, se trouve dans la note 6. Toujours dans la figure ci-dessus, on constante que les candidats d'âge moyenne les plus élevés ont eu la note 10 pour les deux sexes. Et on observe aussi que le sexe garçon a toujours l'âges moyennes les plus élevés que le sexe fille. Cette situation se trouve à partir du note 2 jusqu'à la note 7. À partir de la note 8, 9 et 10, la situation est inversée, c'est-à-dire, c'est le sexe fille qui est le plus âgé en termes de l'âge moyenne.

Après avoir fait l'analyse général de tous les candidats, nous allons découvrir les candidats qui ont eu la moyenne, la note 5 et/ou plus. A noté que l'information sur le nombre des candidats retenus de cet entretien n'a pas été communiquer, nous fixons notre deuxième partie d'analyse sur la note moyenne et la note la plus élevée, par sexe, parcours et niveau.

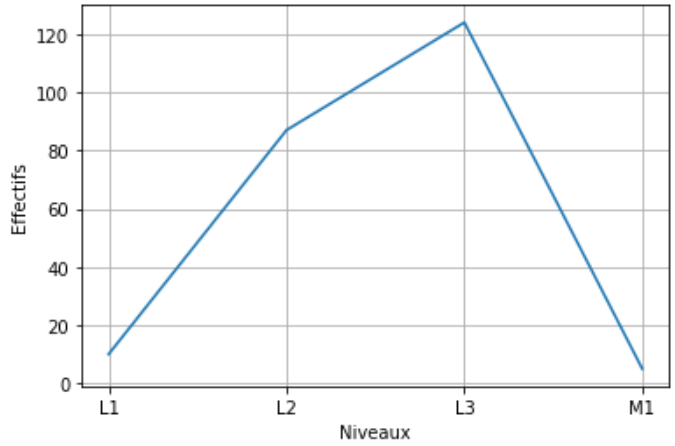
Concernant les candidats qui ont eu la moyenne, la note 5 et/ou plus, sont au nombre de 226 candidats qui représente 43,54% d'effectif total, dont 140 filles et 86 garçons. Parmi les 226, 54,86% sont en L2 et M1 représente que de 2,21% : 10 en L1, 87 en L2, 124 en L3 et 5 en M1.

```
Out[14]: F    140
         G     86
         Name: sexe, dtype: int64

Entrée [15]: entretien_dp[entretien_dp[

Out[15]: L3    124
         L2     87
         L1     10
         M1      5
         Name: niv, dtype: int64
```

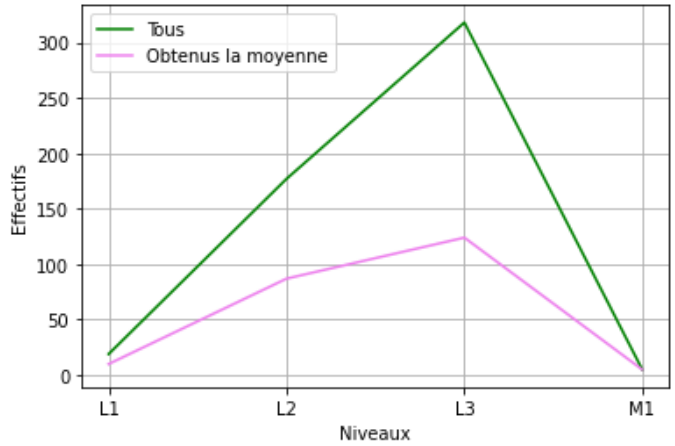
Pour la réparation des candidats qui ont la note 10 et supérieur, selon le résultat ci-dessus, nous pouvons apercevoir dans la figure suivante :



Comme la figure ci-dessus nous montre, la courbe monte à partir du niveau L1 jusqu'au niveau L3. La courbe atteint le pic au niveau L3 car le niveau L3 est le plus nombre. Une descente après jusqu'au M1 qi est notre dernier niveau dans cette analyse du résultat d'entretien.

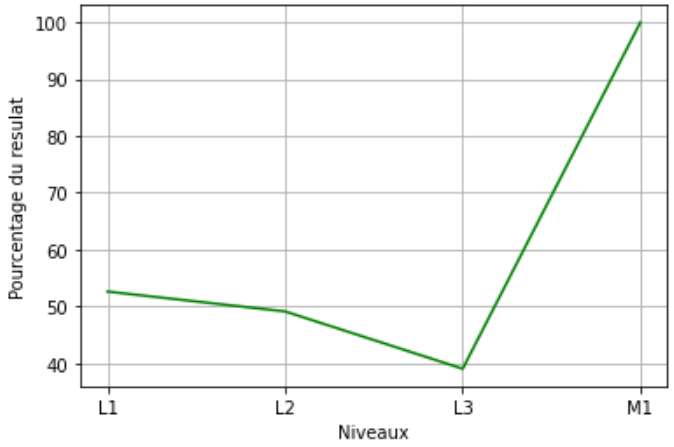
Pour chaque parcours, il y a des candidats qui ont la note moyenne 5 sur 10 ou/et plus, malgré leurs différents effectifs. Des différents sexes dans chaque parcours s'y trouvent sauf dans le parcours Environnement, Géographie, Histoire et Médecine qui sont que de sexe féminin. Et dans les parcours Gestion, Informatique et Tourisme se trouvent que de sexe masculin.

En fait le rapport entre l'effectif total par chaque niveau avec l'effectif des candidats qui de la note moyenne par niveau, nous obtiendrons la figure suivante :



On constate que les deux courbes s'éloignent jusqu'au niveau L3 et se joignent au niveau M1. Cela signifie quel le rapport entre les effectifs se diminue jusqu'à niveau L3 et il prend plus de valeur égalité au niveau M1. Les niveaux L3 est le moins nombreux par rapport à leurs effectifs, qui n'a pas eu la moyenne, alors que les niveau M1 ont tous eu la moyenne. Les L1 ont dépassé la moitié de leurs effectifs. Les L2 sont en dessous de leurs effectifs.

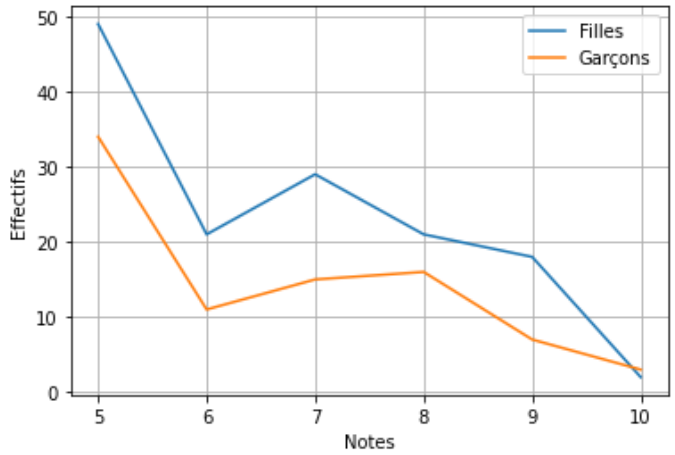
En élargissant notre analyse, on groupe les candidats qui ont eu la note moyenne par parcours, le niveau et avec la prise en compte du sexe. Les niveau L3 en Géographie sont le plus nombreux. Ils sont au nombre de 16, qui ont réussi l'entretien d'embauche. Les L3 en Économie, en Informatique et en Médecine ont 15 candidats de chaque. Et un candidat de chaque pour les parcours et le niveau : L1 en Environnement, M1 en Géographie, M1 en Histoire, L1 en Informatique, M1 en Informatique, L1 en Médecine et L1 en Tourisme.



parcours	niv	sexe	
DROIT	L2	G	8
		F	2
	L3	G	7
ECONOMIE	L1	F	2
	L2	F	5
		G	1
	L3	F	15
ENVIRONNEMENT	L1	F	1
	L2	F	11
	L3	F	9
GEOGRAPHIE	L1	F	2
	L2	F	5
	L3	F	16
	M1	F	1
GESTION	L1	G	2
	L2	G	8
	L3	G	13
HISTOIRE	L2	F	8
	L3	F	13
	M1	F	1
INFORMATIQUE	L1	G	1
	L2	G	7
	L3	G	15
	M1	G	1
MATHS	L2	F	12
		G	5
	L3	F	13
	M1	F	2
MEDECINE	L1	F	1
	L2	F	6
	L3	F	15
TOURISME	L1	G	1
	L2	G	9
	L3	G	8

Name: sexe, dtype: int64

En groupant par sexe et par note, nous obtenons du résultat suivant :



sexe	note	
F	5	49
	7	29
	6	21
	8	21
	9	18
	10	2
G	5	34
	8	16
	7	15
	6	11
	9	7
	10	3

Les candidats qui ont la note 5, sont beaucoup nombreux les candidats qui ont eu des autres notes supérieures à la moyenne. Sur les candidats ont eu la moyenne, ils représentent plus de 36% alors que 14% ont eu 6 ; 19% ont eu 7 ; 16% ont eu 8 ; 11% ont eu 9 et seulement 2% qui ont eu 10.

Sur les deux sexes, c'est toujours le sexe féminin qui a d'effectif plus élevé sur la note sauf sur la note maximale. Bien que l'effectif du sexe féminin soit déjà plus nombreux que du sexe masculin, mais le rapport entre l'effectif total par sexe et l'effectif total, par sexe, des candidats qui ont eu la moyenne, les candidats féminins ont du meilleur résultat par rapport aux candidats masculins : 140 sur 317 candidates féminines, soit 44% et 86 sur 202 candidats masculins (42%).

Pour notre dernière analyse, nous allons voir les majors des candidats pour l'entretien effectué. C'est-à-dire, ceux qu'ils ont eu la note maximale, égale à 10.

Les candidats qui ont obtenus la note maximale sont :

Out[18]:

	sexe	parcours	age	niv	note
439	G	DROIT	27	L3	10
461	G	INFORMATIQUE	26	L3	10
471	G	GESTION	23	L3	10
485	F	MEDECINE	23	L3	10
518	F	HISTOIRE	34	M1	10

```
Entrée [19]: entretien_dp[entretien_dp['note'] == 1
```

```
Out[19]: G    3
         F    2
         Name: sexe, dtype: int64
```

La note maximale se partage avec trois garçons et deux filles, de parcours différents à savoir : Droit, Informatique, Gestion, Médecine et Histoire. Les parcours Économie, Environnement, Géographie, Maths et Tourisme ne se trouvent seulement parmi qui ont la moyenne, ils absentent dans la liste des candidats qui ont eu la note maximale égale à 10. L'âge minimal et maximal sont de 23 ans et 34 ans. Ils se trouvent tous dans les deux niveaux le plus élevé, en L3 et en M1.

V. Conclusion

Pour conclure, l'analyse des données est une science qui permet d'accompagner les entreprises dans l'amélioration de la recherche et du développement. L'exploitation de données dans l'entreprise requit des experts dans le domaine de l'intelligence artificiel. C'est un travail qui devrait être exécuté et exploité d'une manière scientifique. L'interprétation des résultats ne seront pas toujours évident pour tout le monde. C'est une

tâche devra être toujours à confier à l'expert de Data science. Grâce à l'outil d'analyse et notamment à des nombreuses librairies de Python, l'utilisateur de l'analyse des données peut désormais se consacrer aux tâches les plus complexes à savoir, l'utilisation du *Big Data* et la présentation des résultats.

VI. Référence

[1] Analyse des données en Python : une approche étape par étape | Talent Garden (https://talentgarden.org/fr/data/data-analysis-in-python-a-step-by-step-approach/)

[2] Apprenez pandas, Stack Overflow Documentation, 172 pages (https://riptutorial.com/fr/home)

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

[4] Data mining, Mohamed NEMICHE, Faculté des Sciences d'Agadir,Master MASI, 74 pages.