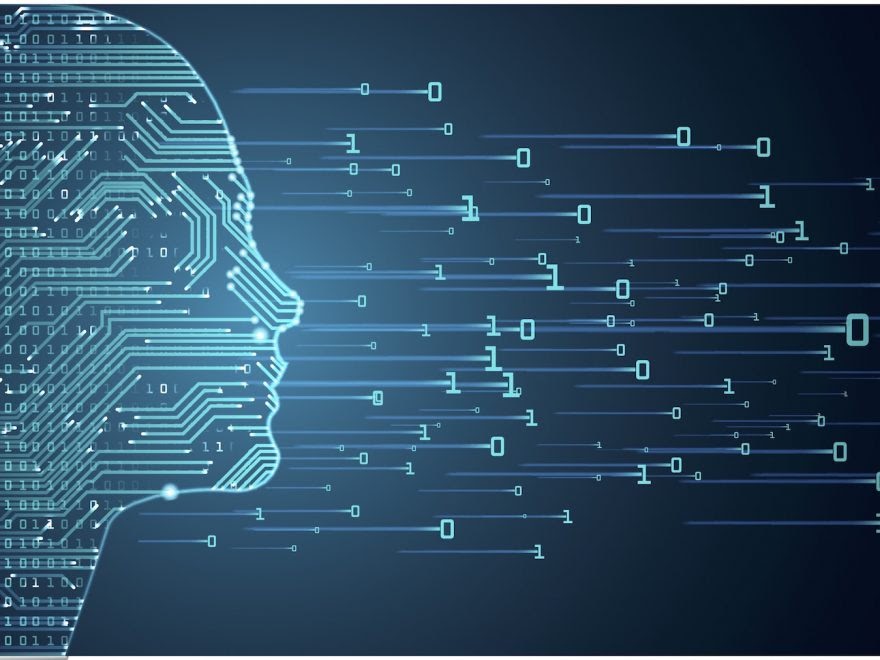
**Rapport des résultats du projet de Data Science**

Sujet : ACP : bilan des entreprises françaises en 2013.

# **Table des matières**

Table des matières

[**Table des matières** 2](#_Toc27862517)

[**Introduction** 3](#_Toc27862518)

[**I)** **Explication du sujet** 4](#_Toc27862519)

[**II)** **Explication technique** 4](#_Toc27862520)

[**III)** **Résultats obtenus et analyses** 4](#_Toc27862521)

[**IV)** **Bilan et conclusion** 4](#_Toc27862522)

# **Introduction**

Le sujet de data science que nous avons choisi est de faire une Analyse en Composante Principales du bilan des entreprise française en 2013. Notre groupe est constitué de Sullivan Honnet et de Jules Vittone, le travail a été organisé en journée de travail, lorsque l’on se réunissait pour travailler on définissait les axes de travail à réaliser durant la séance et on travaillait ensemble.

* Jules a principalement fait de l’élaboration de code ainsi de la recherche pour nous aider à traiter du sujet.
* Sullivan a principalement réalisé le rapport ainsi que du test et de l’analyse des résultats.

    Dans ce rapport nous allons présenter dans un premier temps le sujet que nous avons choisi, puis expliquer les choix techniques que nous avons mis en oeuvre pour réaliser notre analyse qui seront présentés par la suite pour enfin faire un bilan des données.

# **Explication du sujet**

Le but de notre sujet est de réaliser une Analyse en Composantes Principales, ACP, sur un jeu de données contenant les informations relatives aux entreprises français en 2013. Les données ont été récupérées à cette adresse :  <http://www.insee.fr/fr/themes/detail.asp?reg_id=0&ref_id=esane-2013>, qui contient plusieurs types de données pour faire plus d’analyse.

L'intérêt de cette étude est d’en apprendre plus sur les différents liens entre les 25 caractéristiques comptable qui permettent de faire des analyses financières d’une entreprise.

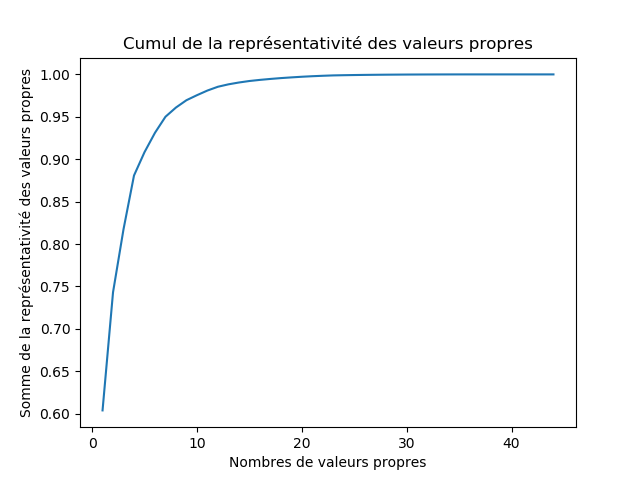
Les caractéristiques sont :

* (a)Nul Nombre d'unités légales
* (b) Cna Capital souscrit non appelé
* (c) Imi Immobilisations incorporelles
* (d) Imc Immobilisations corporelles
* (e) Ter Terrains
* (f) Con Constructions
* (g) Itm Installations techniques, matériel et outillage industriels
* (h) Aic Autres immobilisations corporelles
* Mdt dont matériel de transport
* (j) Ime Immobilisations en cours
* (k) AeA Avances et acomptes
* (l) Imf Immobilisations ﬁnancières
* (m) Tai Total de l’actif immobilisé
* (n) Smp Stocks - Matières premières approvisionnement et en cours
* (o) Sdm Stocks de marchandises
* (p) Aav Avances et acomptes versés sur commandes
* (q) Ccr Clients et comptes rattachés
* (r) Acr Autres créances
* (s) Vmp Valeurs mobilières de placement
* (t) Dis Disponibilité
* (u) Cdr Comptes de régularisation - Charges constatées d’avances
* (v) Tac Total de l’actif circulant
* (w) Acr Autres comptes de régularisation
* (x) Tab Total actif brut
* (y) Tan Total de l’actif net des amortissements et provisions inscrits à l’actif.

# **Explication technique**

Pour ce projet nous avons décidé d’utiliser python pour l’extraction et le traitement des données récupérées, les raisons qui font que nous avons choisi ce langage est principalement car nous maîtrisons le langage, nous avons aussi trouvé de nombreux tutoriaux qui traitent de l’ACP en python.

Nous avons commencé par extraire du fichier les noms des différents secteurs, des caractéristiques et les valeurs associées, ensuite, nous avons supprimé toutes les lignes incomplètes des fichiers et toutes les lignes en double. En entrée, nous avions un fichier d’environ 2000 lignes, en sortie, le tableau obtenu en faisait plus que 800 lignes.

Ce tableau, nous lui faisons subir des transformations en fonction du traitement que l’on souhaite ou des valeurs que nous voulons.

Nous avons fini par inclure dans notre étude les valeurs du passif car nous pensons qu’étudier uniquement l’actif ne permet pas d’avoir des résultats vraiment significatifs. Nous avons donc créé des fichiers contenant toutes les données du passif et nous les avons ajoutées au tableau mentionné ci-dessus. Nous obtenons donc 600 lignes valides à la fin de ce traitement ce qui nous a semblé suffisant pour pouvoir obtenir des résultats concluants.

Par la suite nous devions décider du nombre de caractéristique à prendre en compte pour notre analyse, nous avons donc réalisé un éboulis des valeurs propres et un graphique du cumul de variance restituée selon le nombre de facteurs pour faire cette décision, ensuite nous avons utilisé la règle du coude pour identifier le nombre de facteurs K\* à retenir.

# **Résultats obtenus et analyses**

# **Bilan et conclusion**