



Projet de stage de fin de la deuxième année

Classification des transporteurs par niveau de risque en utilisant k-means clustering

Filière ingénierie e-logistique

Réalisé par :
JARNI Mohamed

Encadré par :
Pr. MAKTITE Mohammed

2019-2020

Remerciements

Je souhaite adresser mes remerciements les plus sincères aux personnes qui m'ont apporté leur aide et qui ont contribué à la réalisation de ce stage ainsi qu'à la réussite de cette formidable année universitaire. Je ne pourrais commencer ce rapport sans présenter mes remerciements les plus loyales à :

Mr. Mohammed MAKTITE pour son accueil, pour avoir accepté la charge de nous encadrer et nous donner une grande liberté d'initiative..

ANP pour son accueil durant cette période de stage, pour tout ce que l'équipe m'a apportée en terme de développement personnel.

Aux membres du jury d'avoir accepté d'évaluer ce travail. Je voudrais aussi exprimer ma vive reconnaissance envers tous les enseignants et le personnel de **L'ECOLE NATIONALE SUPERIEURE D'INFORMATIQUE ET D'ANALYSE DES SYSTEMES** , ainsi que tous ceux qui ont participé à ma formation.

Je n'oublie pas **mes parents** pour leur contribution, leur soutien et leur patience. Enfin j'adresse mes plus sincères remerciements à tous mes proches et amis, qui m'ont toujours soutenue et encouragée au cours de la réalisation de ce projet.

Table des matières

Remerciements	2
Table des figures	5
Introduction	6
I Cadre général du stage	7
1 Présentation de l'organisme d'accueil	7
1.1 Histoire : Naissance de l'ANP	7
1.2 Fiche technique	8
1.3 Services Offertes par l'ANP	8
1.4 Organigramme	9
1.5 Activités et missions des différentes divisions	9
1.5.1 Département police, sûreté, sécurité, environnement	9
1.5.2 Division infrastructures et exploitation régionale	10
1.5.3 Division régulation et développement des places portuaires	10
1.5.4 Division Support	10
2 Description de la mission du stage	11
3 Etude des besoins	13
4 besoins non fonctionnels	13
5 Planification du projet	14
II Analyse exploratoire du fichier reçu	15
6 Technologies utilisées	15
6.1 Numpy	15
6.2 Matplotlib	16
6.3 Pandas	16
7 Poids à vide minimal observé et analyse de la distribution	17
8 Poids chargé minimal observé et analyse de la distribution	17
9 Classification des transporteurs par écart type	18
10 Temps d'attente entre le poids à vide et poids totalement chargé à la sortie	18

11 Destination des transporteurs	19
12 Produit transporté	19
13 Pesage chez different pont bascule	20
 III Classification des transporteurs par niveau de risque en utilisant k-means clustering	 22
14 L'apprentissage non supervisé	22
15 k-means clustering	24
15.1 definition de l'algorithme	24
15.2 La methode du coude	25
15.3 Application de l'algorithme	26
 Conclusion	 28
Bibliographie	29

Table des figures

1	Séparation de l'ODEP	7
2	Organigramme de l'ANP	9
3	Python	15
4	Numpy	16
5	Matplotlib	16
6	Pandas	16
7	distribution poids à vide	17
8	distribution poids chargé	17
9	Poids à vide en catégorie	18
10	Destination des transporteurs	19
11	Catégorie de produits	19
12	Relation entre Poids chargé et Poids à vide par rapport aux catégories de produits	20
13	relation pont poids chargé	21
14	différence entre l'apprentissage supervisé et non supervisé	22
15	exemple d'apprentissage non supervisé	22
16	clustering	23
17	exemple simple de k-means	24
18	methode du coude	26
19	résultat	26

Introduction

Nous vivons aujourd'hui une révolution technologique qui transforme nos vies mais aussi tous les métiers. L'impact de l'intelligence artificielle sur le monde du travail et notre quotidien redistribue le rôle de chacun. De nombreuses entreprises utilisent l'intelligence artificielle dans différents aspects et pour fournir un service à forte valeur ajoutée dans un univers professionnel concurrentiel.

L'objectif de ce stage, comme M. MAKTITE me l'a proposé, est l'application d'algorithme d'apprentissage non supervisé pour classer les transporteurs par niveau de risque, afin de proposer des contrôles détaillés sur les transporteurs à risque.

Ce rapport décline les différentes phases que j'ai suivi pour la réalisation du projet. Il est ainsi organisé en trois chapitres. Le premier chapitre aborde le contexte général du projet. Ensuite, il identifie la problématique et les objectifs du projet ainsi que l'approche adoptée pour sa conduite. Le deuxième chapitre se focalise sur l'analyse du fichier reçu de la part du partenaire . Il explicite les différentes variables ainsi que leurs impact sur la solution finale. Le troisième chapitre est consacré à l'application de l'algorithme k-means qui va nous permettre de classer les transporteurs par niveau de risque afin de proposer des contrôles détaillés .

Première partie

Cadre général du stage

Ce chapitre a pour objet de décrire l'organisation de la société d'accueil ANP, ainsi que les missions envisagées à accomplir durant mon stage.

1 Présentation de l'organisme d'accueil

L'Agence Nationale des Ports est un « Etablissement Public doté de la personnalité morale et de l'autonomie financière ». La tutelle technique de l'Agence est assurée par le Ministère de l'Équipement et du Transport. L'Agence est soumise au contrôle financier de l'Etat applicable aux établissements publics conformément à la législation en vigueur.

1.1 Histoire : Naissance de l'ANP

Promulguée le 15 décembre 2005, la réforme 15-02 relatif au secteur portuaire, a pour l'objet de cesser le monopole commercial de l'Etat, exercé à travers l'Office D'Exploitation des Ports (ODEP) dans le secteur portuaire, ce qui va marquer l'esprit d'ouverture de concurrence aux ports marocains. Depuis les raisons principales est que 95 % des échanges extérieurs du Maroc transitent par voie maritime Croissance économique et compétitivité; la faiblesse du secteur au niveau d'infrastructure (Procédures complexes lors de l'échange, Equipements mal entretenue...)

La mise en œuvre de la réforme : La nouvelle organisation portuaire du Maroc a été entrée en vigueur en Décembre 2006. La mise en œuvre de cette réforme est traduite par : La baisse des tarifs de manutention des conteneurs de 30 %; L'introduction d'un 2ème opérateur privé au port de Casablanca (SOMAPORT); La séparation de l'Office D'Exploitation des Ports (ODEP) en deux entités : - La Société D'Exploitation des Ports (SODEP : Marsa Maroc actuellement), chargée des prestations commerciales. - L'Agence Nationale des Ports (ANP) en charge de l'autorité (capitainerie, régulation ...).

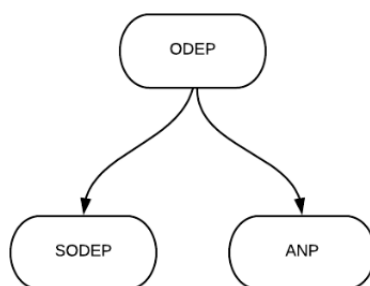


FIGURE 1 – Séparation de l'ODEP

1.2 Fiche technique


Raison sociale :	Agence nationale des ports
Création	28 Décembre 2006 par la loi 15-02
Fondateurs	État marocain
Nom de marque	ANP
Statut	Etablissement Public doté de la personnalité morale et de l'autonomie financière
Président Directoire	DHEM Abdelhakim
Activité	Fédère l'ensemble de la communauté portuaire autour d'objectifs communs
Capital Social	4.062 millions DH
Chiffre d'Affaires	1780.5 millions de DHS
Effectif	2 200
Trafic portuaire	85 millions de tonnes.
Ports exploités	33 Ports (l'ensemble des ports du Royaume à l'exception du port de Tanger Méditerranée)
Téléphone	05 20 12 13 14
Fax	05 22 23 23 35
Site web	www.anp.org.ma
Logo	

TABLE 1 – fiche technique ANP

1.3 Services Offertes par l'ANP

- **Développer les places portuaires :**

- Définir une stratégie de développement de marketing des places portuaires avec et pour l'hinterland (animation de la communauté portuaire) et avec les ports étrangers.

- Optimiser l'organisation de chaque port et définir leur plan d'aménagement.

- Lancer et piloter les projets liés à des potentiels d'exploitation : Extensions de quai, reconversion de Tanger, etc.

- **Réguler l'activité portuaire et gérer les concessions, autorisations et OTDPP :**

- Observer la compétitivité des ports et contrôle l'atteinte des objectifs par les concessionnaires.

- Réguler l'activité des ports sous sa responsabilité en assurant l'observation des règles de la concurrence et de la réglementation (exploitation, police des ports, réglementation sécurité / environnement...), en encadrant les tarifs et en élaborant les cahiers des de charges principaux.

● **Observer la libre concurrence et le bon respect des engagements des concessions et autorisations.**

● **Suivre et gérer les concessions, autorisations et OTDPP :**

- Lancer et piloter les processus de concession ou d'autorisation d'activités commerciales par appel d'offre.

- Gérer la relation avec les concessionnaires et autres partenaires (autorisations).

● **Assurer la sûreté, la sécurité et la police des ports.**

● **Assurer la disponibilité des infrastructures concédées ou non et la continuité de service public (exploitation portuaire) des ports non concédés.**

1.4 Organigramme

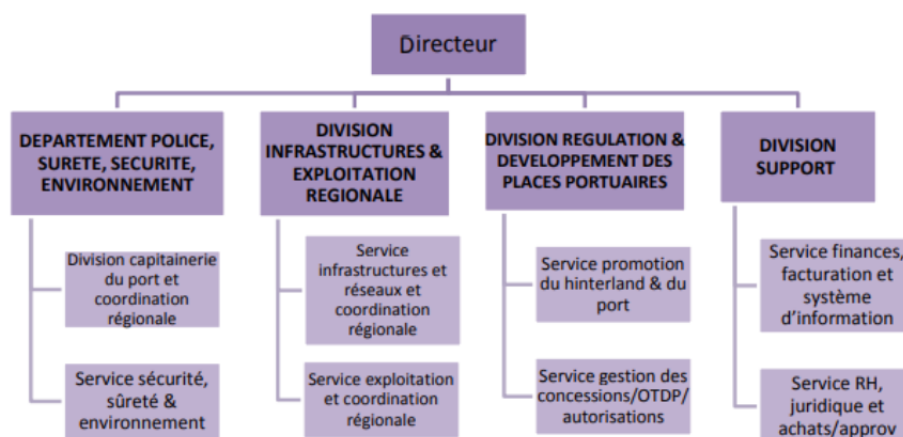


FIGURE 2 – Organigramme de l'ANP

1.5 Activités et missions des différentes divisions

1.5.1 Département police, sûreté, sécurité, environnement

- Prendre en charge le contrôle de l'application de l'ensemble des règles et dispositions réglementaires, régissant les conditions de fonctionnement d'un port.

- Réaliser efficacement des arbitrages nécessaires.

- Développer une politique commune et coordonnée de pilotage et de prévention des risques sur les trois domaines : sûreté, sécurité, protection de l'environnement.
- S'assurer de la conformité des décisions prises avec la réglementation en vigueur, en matière de police portuaire.
 - Assurer la coordination du comité local de la sûreté.
 - Veiller à la conformité du port au code ISPS.

1.5.2 Division infrastructures et exploitation régionale

- Maintenir en bon état les infrastructures portuaires et superstructures mis à la disposition de la Direction Régionale.
 - Assurer la fourniture d'eau, d'électricité aux usagers du port.
 - Assurer la gestion des réseaux d'assainissement du port.
 - Garantir l'hygiène, la salubrité et la protection de l'environnement dans les zones non concédées.

1.5.3 Division régulation et développement des places portuaires

- Animer et élaborer la stratégie par la Région avec l'ensemble des ports de son périmètre.
- Réaliser l'observatoire de compétitivité au niveau régional.
- Suivre et gérer les concessions, les autorisations des activités connexes et d'OTDP.
- Observer la compétitivité des intervenants au sein du port.

1.5.4 Division Support

- Assurer l'encadrement, l'assistance et la logistique pour la Direction.
- Veiller au maintien des équilibres financiers du port.
- Veiller au respect des procédures en matière d'appels d'offres et de consultations.
- S'assurer du respect des procédures en vigueur à l'ANP.
- Contribuer dans la mise en place et la mise à jour des procédures de la direction.
- Satisfaire les besoins des entités de la Direction en matériels bureautique, informatique, consommable, etc.
 - Veiller au respect et à l'harmonisation des procédures approvisionnements et de gestion des stocks.
 - Veiller à l'optimisation de la gestion des stocks.
 - Assurer la mise en place du plan d'équipement et de maintenance microinformatique.

- Diffusion du tableau de bord mensuel des ressources humaines.
- Assurer dans les meilleures conditions de délai, de coût et de sécurité la gestion du parc du matériel informatique et bureautique, la mise en place et la maintenance des applications informatiques.
- L'élaboration et formulation des besoins en matériel informatique, téléphonique et bureautique.

2 Description de la mission du stage

Je suis en charge d'analyser un fichier source, reçu de la part d'un partenaire, dans l'objectif de détecter les potentiels fraudeurs et de s'assurer que le transport de marchandise d'un point d'escale jusqu'au client final s'est passé sans moindre souci.

Le processus de chargement de la marchandise par le transporteur se déroule selon les étapes suivantes :



Saisie des données:

- Num_escale
- Num_vehicule
- Num_remorque
- num_Client
- Date_entrée et Heure_entrée
- Pont_entrée
- Poids_vide



Pesage du camion à vide au pont bascule (Pont_entrée)



Arrivée du transporteur au port



Transporteur se présente au niveau du trémie pour chargement de marchandise



Pesage du camion chargé au pont bascule (Pont_sortie) et saisie de données:

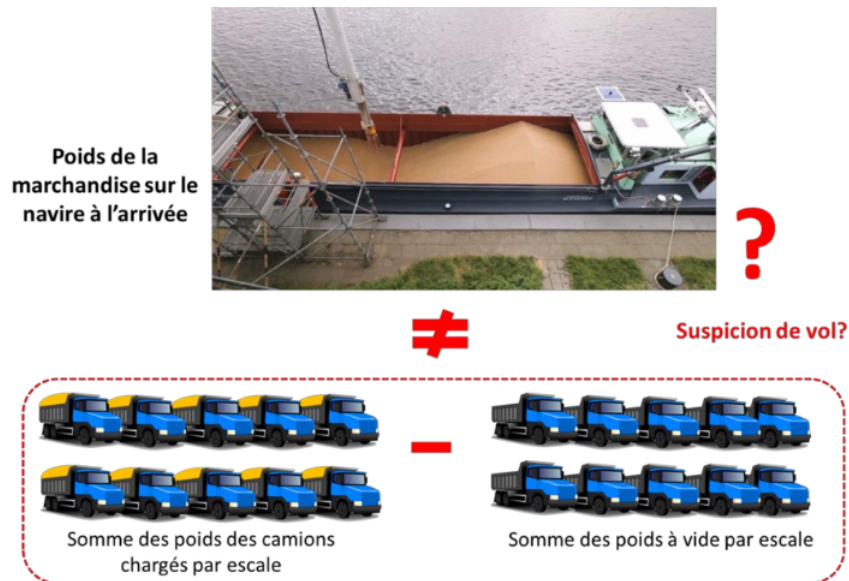
- Pont_Sortie
- Poids chargé
- Produit
- Date_sortie et Heure_sortie



Sortie du transporteur du port avec la marchandise, à destination du client

Le contrôle concerne particulièrement deux étapes : le numéro 2 et le numéro 5.

Lors du processus d'audit habituel, un contrôle en lot a initié . Pour se faire, ils ont vérifié le poids total de la marchandise en provenance de l'import avec la somme de la marchandise chargée par camion. Il a constaté un écart important.



Afin de mener une enquête approfondie, on a demandé les données saisies par le partenaire spécialisé dans l'exploitation des terminaux et des quais portuaires. En retour, on a reçu un fichier contenant les champs suivant :

- Num-escale : Numéro du navire chargé de marchandise
- Num-véhicule : Numéro du transporteur
- Num-Remorque : Numéro de la remorque accrochée au véhicule
- Num-client : Numéro du client
- Date-entrée : La date de l'entrée du transporteur au pont bascule
- Heure-entrée : L'heure de la prise du poids (à vide) au niveau du pont bascule
- Pont-entrée : Numéro du pont à l'entrée
- Pont-sortie : Numéro du pont de bascule pour la prise du poids du camion chargée
- Date-sortie : La date de la prise du poids (chargé) au niveau du pont
- Heure-sortie : L'heure de la prise du poids (chargé) au niveau du pont
- Poids-Vide : Le poids du camion vide
- Poids-chargé : Le poids du camion chargé de la marchandise
- Produit : La marchandise

Liste de points de vigilance :

- Le camion ne s'est pas correctement positionné
- Le transporteur se présente à un pont bascule pour la mesure de poids à vide et change de pont pour la mesure du poids chargé
- Le transporteur cache des objets lourds lors du premier pesage à vide et se débarrasse du poids ou objets sur le chemin de son chargement.

3 Etude des besoins

En terme de besoins, le stage avait l'objectif d'analyser le fichier reçu afin de mettre en évidence les fraudeurs potentiels. L'analyse se fera selon les axes suivant :

1. Le poids à vide (chauffeur inclut) minimal observé et l'analyse de la distribution.
2. Le poids chargé (chauffeur inclut) minimal observé et l'analyse de la distribution.
3. Classification des transporteurs par écart type :
 - a. Liste rouge :camions venant avec 2 fois l'écart type.
 - b. Liste orange :camions venant avec 1 fois l'écart type.
 - c. Liste verte : camions venant avec + 150 kg de plus que le poids à vide minimal observé.
4. Temps d'attente entre le poids à vide et poids totalement chargé à la sortie.
5. La destination des transporteurs.
6. Le produit transporté.
7. Pesage chez différent pont bascule.

Après avoir terminé la partie d'analyse . Le besoin ensuite était de classer les transporteurs par niveau de risque en faisant appel à l'apprentissage non supervisé qui est une méthode d'apprentissage automatique dans laquelle au lieu de montrer à la machine des exemples (X,y) de ce qu'elle doit apprendre. On lui fournit uniquement des données X et on lui demande d'analyser la structure de ces données pour apprendre elle même à réaliser certaines tâches.

4 besoins non fonctionnels

Les besoins non fonctionnels représentent toutes les contraintes auxquelles est soumis l'algorithme pour son application et son bon fonctionnement.

- Choix du modèle : algorithme de clustering utilisé avec tous ses paramètres.
- Performance : Le temps d'exécution ne devra pas dépasser plusieurs secondes dans le cas des fonctionnalités complexes.

— Qualité : le graphe des résultats de l'algorithme doit être simple et facile à comprendre et interpréter.

5 Planification du projet

Durant le cycle de la réalisation du stage, j'ai suivi le planning suivant : j'ai commencé par la rédaction du cahier de charge du projet et la spécification des besoins et leurs faisabilités. Ensuite, j'ai démarré la partie analyse du fichier reçu de la part du partenaire. Puis, j'ai commencé la recherche sur l'apprentissage non supervisé, son principe, ses applications et ses différents algorithmes. Enfin, j'ai démarré l'application de l'algorithme choisi et interpréter ses résultats.

Dans ce chapitre j'ai présenté le contexte général du projet et les différents besoins fonctionnels et non fonctionnels du projet ainsi que sa conduite.

Deuxième partie

Analyse exploratoire du fichier reçu

Cette partie du rapport est consacrée aux modalités de l'analyse des différents composants du fichier. Dans un premier sous chapitre, je présente les différentes technologies utilisées. Ensuite, j'entame l'analyse.

6 Technologies utilisées

Le langage utilisé dans la réalisation du projet est Python, qui techniquement est un langage de programmation, mais qui est devenu grâce à ses nombreux outils la référence en matière de traitement de données. [1]



FIGURE 3 – Python

Python possède plusieurs propriétés qui en font un outil de choix pour l'analyse de données :

Simplicité : c'est un langage de programmation simple. On voit la différence en comparant la complexité d'un programme simple, qui imprime "Hello world", en langage C et en Python.

Interactivité : c'est un langage interprété. Cela signifie en pratique qu'on peut lancer un script en Python sans étape de compilation. On peut écrire et exécuter notre programme ligne par ligne, en examinant à chaque fois le résultat. Cette propriété est exploitée à merveille par les notebooks Colaboratory (que nous utiliserons dans la suite dans le projet).

L'écosystème : Peut-être le point le plus important, est l'ensemble des outils, des bibliothèques, la communauté, bref, tout ce qui gravite autour du langage lui-même. Python possède des bibliothèques pour à peu près tout ce qu'on peut imaginer.

6.1 Numpy

Diminutif de Numerical Python, Numpy fournit une interface pour stocker et effectuer des opérations sur les données. D'une certaine manière, les tableaux Numpy sont comme les listes en Python, mais Numpy permet de rendre les opérations beaucoup plus efficaces, surtout sur

les tableaux de large taille. Les tableaux Numpy sont au cœur de presque tout l'écosystème de data science en Python.[2]



FIGURE 4 – Numpy

6.2 Matplotlib

Matplotlib a vu le jour pour permettre de générer directement des graphiques à partir de Python. Au fil des années, Matplotlib est devenu une librairie puissante, compatible avec beaucoup de plateformes, et capable de générer des graphiques dans beaucoup de formats différents.



FIGURE 5 – Matplotlib

6.3 Pandas

Avec Numpy et Matplotlib, la librairie Pandas fait partie des librairies de base pour la data science en Python. Pandas fournit des structures de données puissantes et simples à utiliser, ainsi que les moyens d'opérer rapidement des opérations sur ces structures.[3]



FIGURE 6 – Pandas

Le dataset contient 70609 ligne et 13 colonne

7 Poids à vide minimal observé et analyse de la distribution

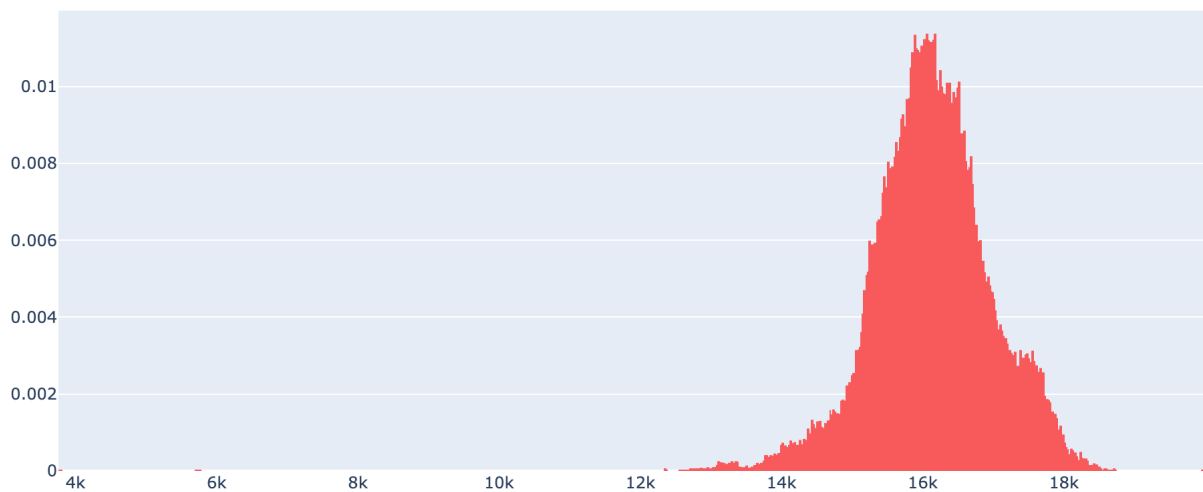


FIGURE 7 – distribution poids à vide

- Le poids à vide minimale observée est 3760 qui est une valeur aberrante.
- Le poids à vide moyen observée est 16120.
- La distribution observée est normale, avec un écart type de 833, la plupart des poids sont entre 15k et 17k.

8 Poids chargé minimal observé et analyse de la distribution

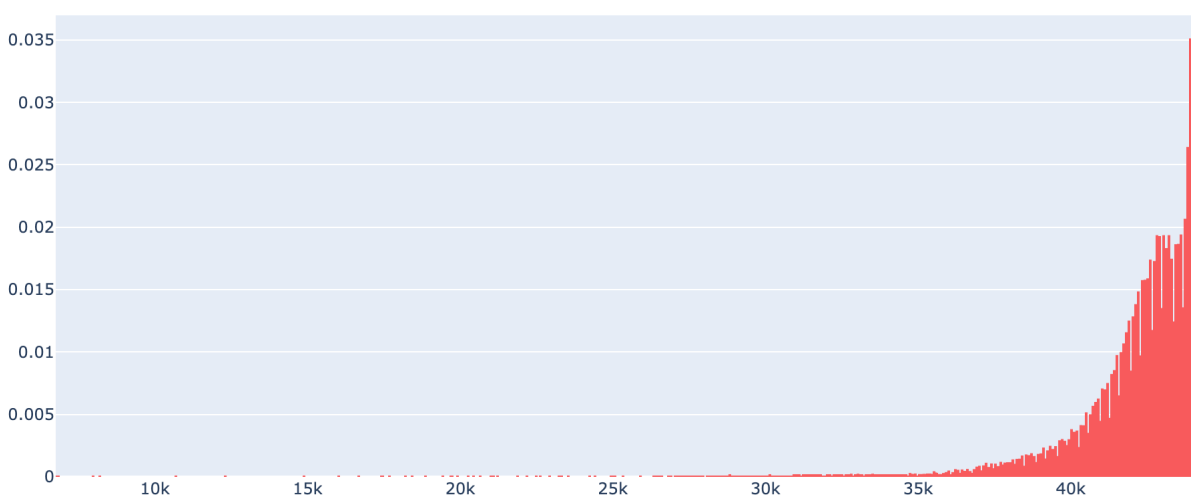


FIGURE 8 – distribution poids chargé

- Le poids chargé minimale observée est 6760 qui est une valeur aberrante

- Le poids chargé moyen observée est 42017
- La distribution observée présente une asymétrie vers la gauche, c'est-à-dire que les grandes valeurs observées sont plus fréquentes que les valeurs plus petites, la plupart des poids sont supérieures à 40k.

9 Classification des transporteurs par écart type

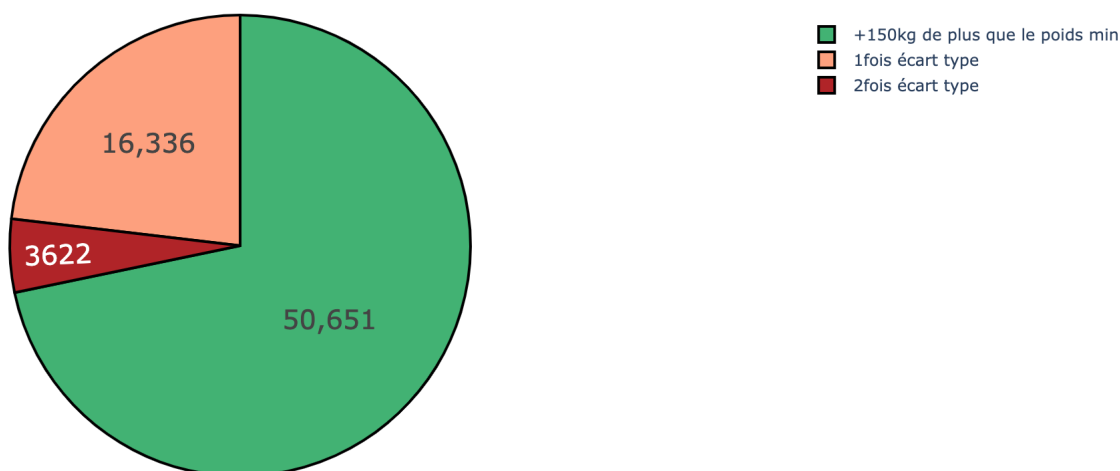


FIGURE 9 – Poids à vide en catégorie

transformation du poids en catégories :

- rouge = camions venant avec 2 fois écart type
- orange = camions venant avec 1 fois écart type
- vert = camions venant avec +150kg de plus que le poids min

pourcentage de chaque catégorie : vert (71.7%) ; orange (23.1%) ; vert (5.2%)

10 Temps d'attente entre le poids à vide et poids totalement chargé à la sortie

Après avoir calculer la difference entre le temps d'arrivée des transporteurs pour le pesage du poids à vide et le temps de sortie en destination du client, on a trouvé que 70277 des cas ne dépassaient pas une journée , ce qui est plutot logique , 109 duraient plus que 24h et un seul cas avec plus de 48h. Par contre la difference était négative pour 222 cas . Peut-être qu'il y avait une erreur dans le saisie ou qu'il y avait une sorte de triche.

11 Destination des transporteurs

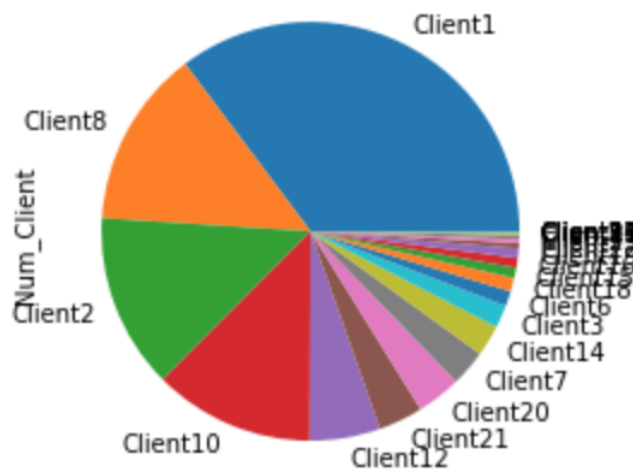


FIGURE 10 – Destination des transporteurs

La plupart des transporteurs sortent du port avec la marchandise à destination du client1 suivi du client8 où la plupart des valeurs du temps d'attente sont négatifs (149/222) suivis du client2.

12 Produit transporté

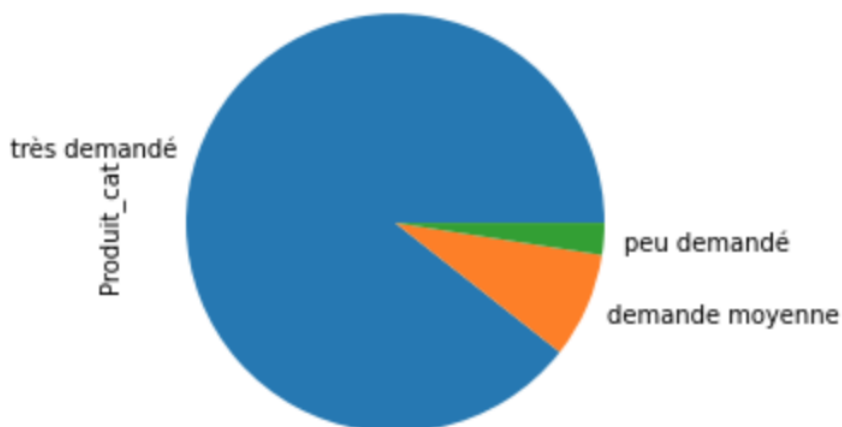


FIGURE 11 – Catégorie de produits

Les produits transportés plus de 1000 fois sont : MAIS, TRTX DE SOJA, BLE TENDRE, TRT DE TOURNESOL, ORGE, BLE DUR, CHARBON, PULPE DE BETTRAVE, SON DE BLE, DDGS, GRAINE DE SOJA, COQUE DE SOJA, TRTX DE COLZA, huile brute de soja.

Les produits transportés entre 100 et 1000 fois sont : COKE DE PETROLE , HUILE DE SOJA, V DDGS, CORN GLUTEN, PNEUS DECHIQUETES, AVOINE, GLUTEN DE

MAIS, VRAC ORGE, ferraille, HUILE BRUTE DE PALME, SABLE, V GLUTEN, VRAC D'ANTHRACITE, mais, TRTX SOJA, TRTX DETOURNESOL', 'huile brute de tournesol ,HUILE BRUTE DE TOURNESOL, KHARROUB CONCASSEES.

Les produits transportés moins de 100 fois sont : huile de soja, ARGILE, ble dur, hydroxyde d aluminium, VRAC D'ARGILE, HUILE BRUTE DE SOJA, S/BLE, TRTX TOURNESOL, lot de frdx steel billets, SEL GEMME, RECUPERATION CORDAGE, RECUPERATION DE FARDAGE, nouveau bois usagée, BOIS DE FARDAGE.

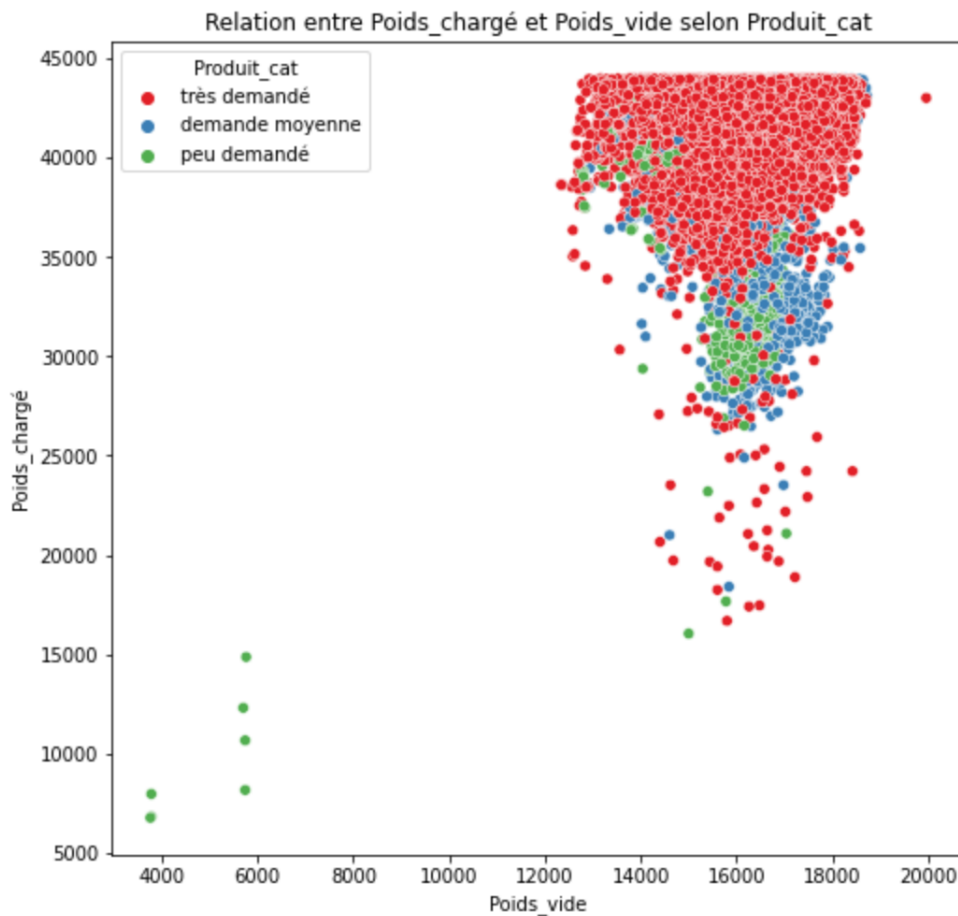


FIGURE 12 – Relation entre Poids chargé et Poids à vide par rapport aux catégories de produits

13 Pesage chez different pont bascule

Entrée par un pont de bascule et sortie par un autre : les camions se presentant dans des ponts differents sont-ils susceptible de frauder ?

nombre de peusage chez différents ponts est 23861 (33%).

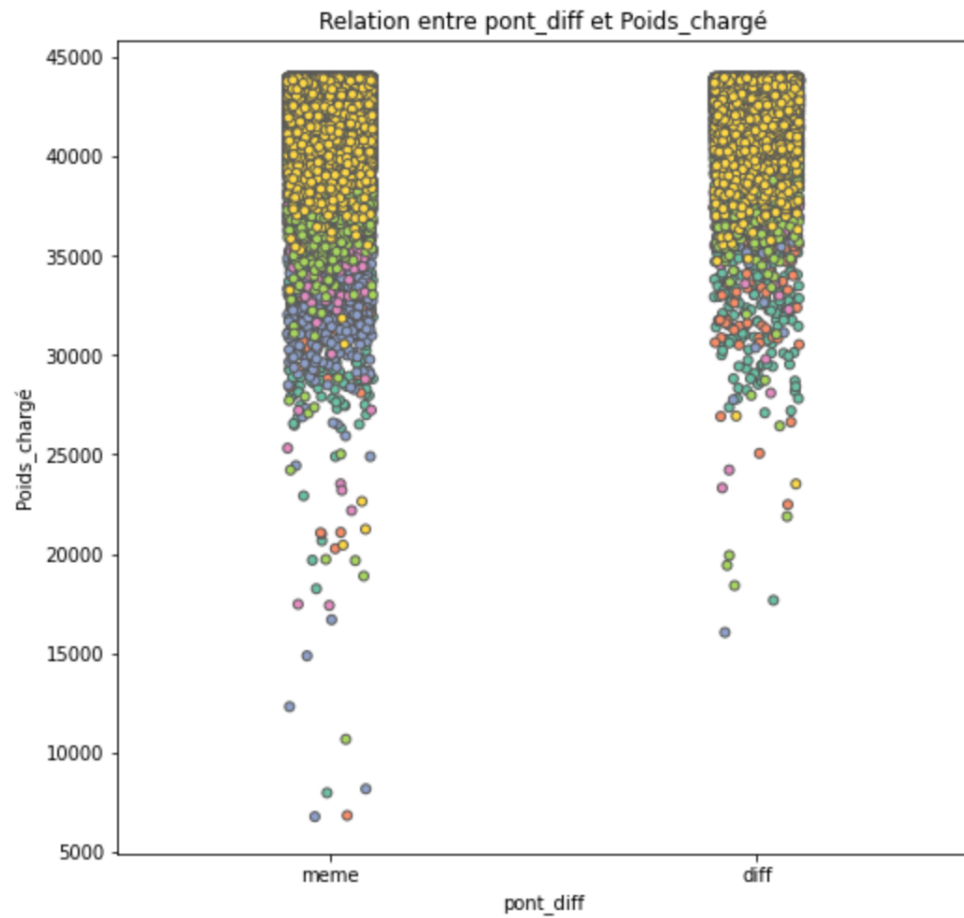


FIGURE 13 – relation pont poids chargé

On remarque que le pont de pesage d'entrée soit le meme que celui de sortie ou bien différent n'influence pas sur le poids chargé.

Troisième partie

Classification des transporteurs par niveau de risque en utilisant k-means clustering

Dans cette partie, je vais appliqué l'algorithme de l'apprentissage non supervisé : k-means pour classer les différents transporteurs par niveau de risque de fraudes.

14 L'apprentissage non supervisé

L'apprentissage non supervisé est une méthode d'apprentissage automatique dans laquelle au lieu de montrer à la machine des exemples (X,y) de ce quelle doit apprendre. On lui fournit uniquement des données X et on lui demande d'analyser la structure ces données pour apprendre elle meme à réaliser certaines taches.



FIGURE 14 – différence entre l'apprentissage supervisé et non supervisé

Par exemple la machine peut apprendre à classer des données en les regroupant uniquement selon leurs ressemblances. C'est ce qu'on appelle faire de la classification non supervisée.



FIGURE 15 – exemple d'apprentissage non supervisé

Voici les algorithmes d'apprentissage non supervisé les plus importants :

- Clustering
 - k-Means
 - Hierarchical Cluster Analysis (HCA)
 - Expectation Maximization
- Visualisation et reduction de dimension
 - Principal Component Analysis (PCA)
 - Kernel PCA
 - Locally-Linear Embedding (LLE)
 - t-distributed Stochastic Neighbor Embedding (t-SNE)
- Association rule learning
 - Apriori
 - Eclat

Celui qui nous interesse c'est le **clustering**. Par exemple, disons que vous avez beaucoup de données sur les visiteurs de votre blog. Vous pouvez exécuter un algorithme de regroupement pour essayer de détecter des groupes de visiteurs similaires . À aucun moment vous ne dites à l'algorithme à quel groupe appartient un visiteur : il trouve ces connexions sans votre aide. Par exemple, il pourrait remarquer que 40 % de vos visiteurs sont des hommes qui aiment les bandes dessinées et qui lisent généralement votre blog le soir, tandis que 20 % sont de jeunes amateurs de science-fiction qui vous rendent visite le week-end, etc. Si vous utilisez un algorithme de clustering, il peut également subdiviser chaque groupe en plus petits groupes. Cela peut vous aider à cibler vos messages pour chaque groupe. [4]

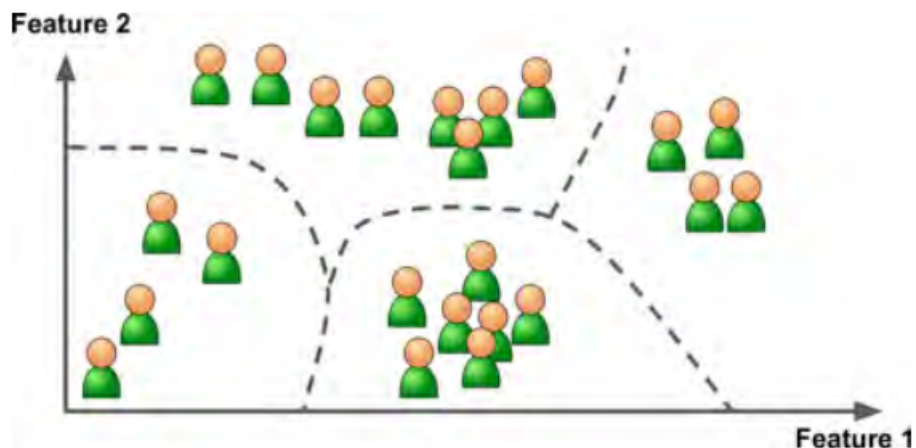


FIGURE 16 – clustering

15 k-means clustering

15.1 definition de l'algorithme

K-means est l'un des algorithmes d'apprentissage non supervisé les plus simples et les plus populaires. Son objectif est simple : regrouper des points de données similaires et découvrir des modèles sous-jacents. Pour atteindre cet objectif, K-means recherche un nombre fixe (k) de **clusters** dans un ensemble de données.

Un **cluster** désigne un ensemble de points de données regroupés en raison de certaines similarités.

Vous définirez un nombre cible k , qui se réfère au nombre de centroïdes dont vous avez besoin dans l'ensemble de données. Un centroïde est l'emplacement imaginaire ou réel représentant le centre du **cluster**. Chaque point de données est attribué à chacun des **clusters** en réduisant la somme des carrés à l'intérieur de la grappe. En d'autres termes, l'algorithme **K-means** identifie un nombre k de centroïdes, puis attribue chaque point de données au **cluster** le plus proche, tout en gardant les centroïdes aussi petits que possible.

La "moyenne" dans **K-means** fait référence à la moyenne des données, c'est-à-dire à la recherche du centroïde.

Pour traiter les données d'apprentissage, l'algorithme **K-means** commence par un premier groupe de centroïdes sélectionnés au hasard, qui sont utilisés comme points de départ pour chaque groupe, puis effectue des calculs itératifs (répétitifs) pour optimiser les positions des centroïdes. Il cesse de créer et d'optimiser les clusters lorsque :

- Les centroïdes se sont stabilisés - il n'y a pas de changement dans leurs valeurs parce que le regroupement a été réussi.

ou :

- Le nombre défini d'itérations a été atteint.[5]

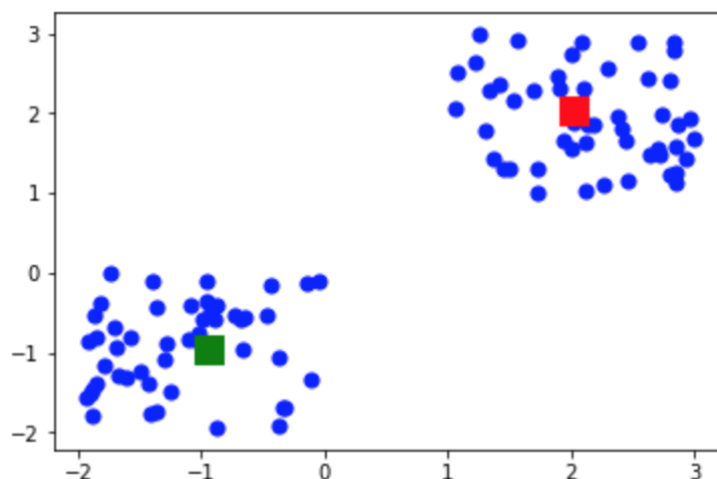


FIGURE 17 – exemple simple de k-means

15.2 La methode du coude

La méthode du coude est principalement utilisée dans les algorithmes d'apprentissage non supervisé pour déterminer le nombre optimal de clusters à utiliser pour trouver des groupes inconnus spécifiques au sein de notre population. Rappelez-vous que dans K-means clustering, nous ajoutons le nombre de clusters de manière manuelle, donc la méthode du coude est utile lorsque vous utilisez K-means.

Pourquoi l'appelle-t-on la méthode du coude ? Parce qu'au fur et à mesure des itérations pour trouver le nombre optimal de groupes, la ligne prend la forme du bras et le nombre optimal de groupes est le point qui se trouve dans la partie du coude du bras.

La distance euclidienne

$$d(x, y) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (1)$$

x_1 = valeur de l'axe des X de l'observation des données.

x_2 = valeur de l'axe X du centroïde du cluster.

y_1 = valeur de l'axe des Y de l'observation des données.

y_2 = valeur de l'axe Y du centroïde du cluster.

La méthode du coude permet de trouver la somme moyenne des carrés de la distance entre le centre du cluster et les observations de données. Plus le nombre de clusters augmente, plus la somme moyenne des carrés diminue. Et, plus le nombre de clusters augmente, plus la distance entre les points de données et les centroïdes diminue également. À chaque fois, nous voyons le "coude" qui est une règle empirique pour considérer le nombre optimal de clusters.

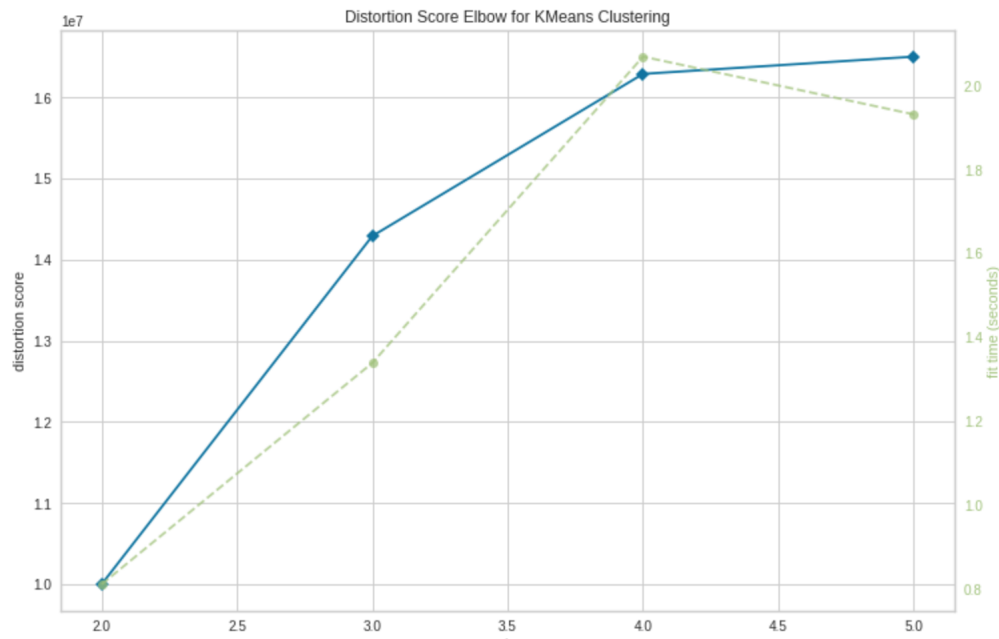


FIGURE 18 – methode du coude

D'après le graphe, le nombre de cluster qu'on va choisir est 4.

15.3 Application de l'algorithme

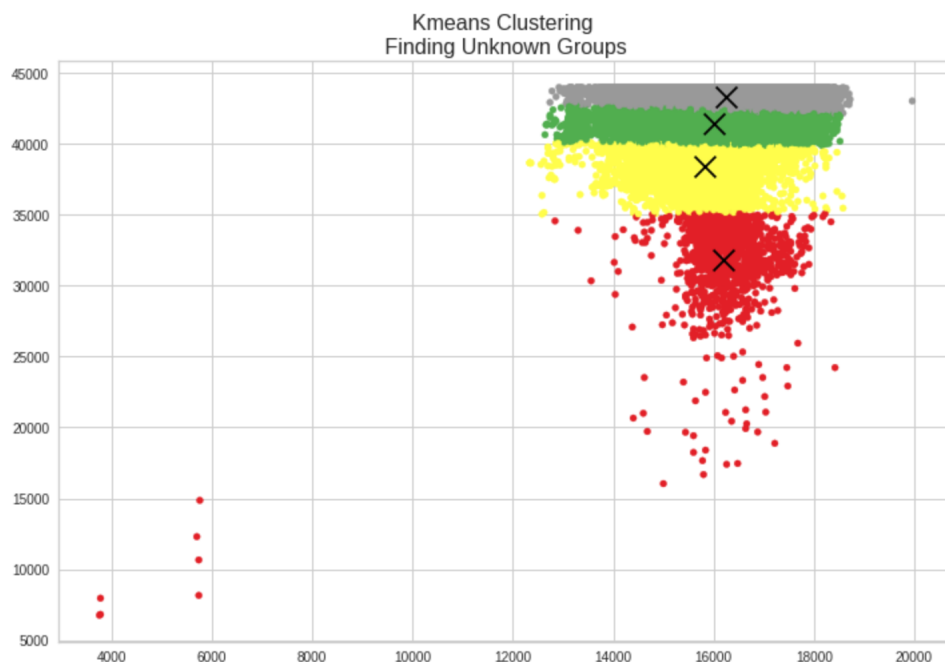


FIGURE 19 – résultat

L'algorithme a divisé la population en 4 parties, celle en rouge présente un gros risque et potentiel de fraude surtout les cas en bas à gauche où la différence entre poids à vide et poids

chargé est assez faible. Donc il faut réaliser des contrôles détaillés sur ses transporteurs. Pour la partie jaune le risque est moins présent comparant à la partie rouge. Par contre les parties verte et grise ne présente aucun risque.

Conclusion

Ce stage était l'incarnation d'une mission purement riche et professionnelle. Le sujet entre dans le thème d'utilisation de l'intelligence artificielle qui est devenue une tendance lors de ces dernières années au sein des entreprises, et portait sur l'application d'un algorithme de machine learning à un fichier reçu ayant pour objectif la classification des transporteurs par niveau de risque.

Pour atteindre cet objectif, j'ai commencé par analyser le fichier reçu de la part du partenaire ce qui m'a permis de mieux comprendre ses différentes composantes et la relation entre eux et avoir une idée beaucoup plus claire sur le cas traité.

J'ai ensuite abordé le principe de l'apprentissage non supervisé, ses algorithmes les plus populaires, k-means clustering en particulier.

Finalement, en utilisant la méthode du coude pour déterminer le nombre de cluster qu'il faut choisir, j'ai attaqué la partie d'application de l'algorithme k-means clustering.

Ce stage m'a été bénéfique sur plusieurs plans. Il m'a permis de maîtriser les méthodes, concepts et outils de l'apprentissage non supervisé, de consolider ma formation théorique et pratique relativement aux phases d'un projet data science et de me familiariser avec les bibliothèques de Python.

Bibliographie

- [1] Ali Neishabouri. Découvrez les librairies python pour la data science. <https://openclassrooms.com/fr/courses/4452741-decouvrez-les-librairies-python-pour-la-data-science/5560976-familiarisez-vous-avec-lecosysteme-python>.
- [2] Ali Neishabouri. Plongez en détail dans la librairie numpy. <https://openclassrooms.com/fr/courses/4452741-decouvrez-les-librairies-python-pour-la-data-science/4740941-plongez-en-detail-dans-la-librairie-numpy>.
- [3] Ali Neishabouri. Passez de numpy à pandas. <https://openclassrooms.com/fr/courses/4452741-decouvrez-les-librairies-python-pour-la-data-science/5558996-passez-de-numpy-a-pandas>.
- [4] Aurélien Géron. Hands-on machine learning with scikit-learn and tensorflow.
- [5] Dr. Michael J. Garbade. Understanding k-means clustering in machine learning. <https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>.