REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
École Nationale Polytechnique

Schlumberger

Département : Génie Industriel
Entreprise : Schlumberger NAF

Mémoire de Projet de Fin d'Études
En vue de l'obtention du diplôme d'Ingénieur d'État en Génie Industriel
Option : Management Industriel

Estimation des lead times liés à l'importation à travers l'apprentissage machine dans le cadre de la méthodologie CRISP-DM

Application : Schlumberger NAF

Mohamed Annis SOUAMES
Larbi Abderrahmane MOHAMMEDI

Sous la direction de M. Iskander ZOUAGHI

Présenté et soutenu publiquement le (27/06/2022)

Composition du jury :

Présidente	Mme. Bahia BOUCHAFAA	MCA	ENP
Examinateur	M. Oussama ARKI	MCB	ENP
Promoteur	M. Iskander ZOUAGHI	MCA	ENP
Invité	Mme. Hadia OUAFI	Algeria Import-Export Leader	Schlumberger

ENP 2022

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
École Nationale Polytechnique

Schlumberger

Département : Génie Industriel
Entreprise : Schlumberger NAF

Mémoire de Projet de Fin d'Études
En vue de l'obtention du diplôme d'Ingénieur d'État en Génie Industriel
Option : Management Industriel

Estimation des lead times liés à l'importation à travers l'apprentissage machine dans le cadre de la méthodologie CRISP-DM

Application : Schlumberger NAF

Mohamed Annis SOUAMES
Larbi Abderrahmane MOHAMMEDI

Sous la direction de M. Iskander ZOUAGHI

Présenté et soutenu publiquement le (27/06/2022)

Composition du jury :

Présidente	Mme. Bahia BOUCHAFAA	MCA	ENP
Examinateur	M. Oussama ARKI	MCB	ENP
Promoteur	M. Iskander ZOUAGHI	MCA	ENP
Invité	Mme. Hadia OUAFI	Algeria Import-Export Leader	Schlumberger

ENP 2022

ملخص

يهدف هذا العمل إلى تقدير المهل الزمنية المتعلقة بعملية الاستيراد داخل شركة الخدمات النفطية Schlumberger NAF**باستخدام خوارزميات مختلفة للتعلم الآلي بالإضافة إلى تقنيات توليد البيانات لمعالجة مشكلة نقصها ، مع اتباع منهجية** CRISP-DM **كإطار مرجعي.**

تتمثل هذه المنهجية أولاً في تحديد المشكلة من وجهة نظر تجارية وتقنية ، وجمع البيانات وتنظيفها من أجل استخدامها في خوارزميات التعلم الآلي المختلفة ، وأخيراً نشر الحل في شكل تطبيق Chat-bot

يساهم حلنا في تخفيض التكاليف بالإضافة إلى توقع المواعيد المتعلقة بالاستيراد في شركة Schlumberger NAF.

الكلمات المفتاحية : الاستيراد، المهل الزمنية، تعلم الالة، توليد البيانات، منهجية CRISP-DM ، Chat-bot.

Abstract

The objective of this work is to estimate the lead times related to the import process within the oil services company Schlumberger NAF using different machine learning algorithms and data synthetisation techniques to generate a larger volume of data, while following the CRISP-DM approach as a reference framework.

The implementation of this approach consists first in defining the problem from a business and a technical perspective, collecting and cleaning the data, and then use it in the different machine learning models and finally deploying the solution through a chat-bot interface.

Our solution contributes to the optimization of costs and the anticipation of import lead times.

Keywords : Imports, Lead times, Machine learning, Data synthetisation, CRISP-DM, Chat-bot.

Résumé

Ce présent travail a pour objectif d'estimer les lead times liés au processus d'importation au sein de l'entreprise de services pétroliers Schlumberger NAF en utilisant les différents algorithmes d'apprentissage machine ainsi que les techniques de génération (synthétisation) des données pour remédier au problème du faible volume de données, tout en suivant l'approche CRISP-DM comme cadre de référence.

La mise en place de cette démarche consiste tout d'abord à définir le problème d'un point de vue métier et technique, collecter les données et les nettoyer afin de les utiliser dans les différents modèles de Machine Learning et enfin déployer la solution sous forme d'une interface de Chat-bot.

Notre solution contribue à l'optimisation des coûts ainsi qu'à l'anticipation des délais d'importation au sein du département import-export de Schlumberger.

Mot clés : Importation, Lead times, Apprentissage machine, Génération des données, CRISP-DM, Chat-bot

Dédicaces

Louange à Dieu tout-puissant, qui m'a permis de voir ce jour tant attendu arriver

Je dédie ce travail à

Ma très chère mère, source inépuisable de tendresse, de patience, qui a fait tant de sacrifices, je voudrais te remercier pour ton amour, ta générosité et ton soutien, que Dieu tout-puissant te protège et t'accorde une longue vie pleine de bonheur, de santé et de prospérité.

Mon très cher père, qui a toujours été un exemple pour moi, qui m'a inculqué le sens du travail et de la responsabilité, je souhaite te remercier pour tes sacrifices, ta compréhension et ton soutien, qu'Allah te préserve pour nous et te procure santé et bonheur.

À ma chère petite sœur, Rayhane, une sœur qu'on ne trouve nulle part ailleurs, je te souhaite une longue vie pleine de réussite et de succès et que Dieu te protège inch'Allah.

À mon cher petit frère, Safouane, que j'aime tellement, je te souhaite plein de réussite dans tes études et dans ta vie, que Dieu te préserve inch'Allah.

À ma chère grand-mère paternelle, "Mama-Oua", qui a toujours fait preuve d'affection, de soutien et de tendresse, j'espère que ce travail te rendra fière et qu'Allah t'accorde une longue vie pleine de santé et de prospérité.

À la mémoire de mon grand-père paternel "Papa-Djedou", mon grand-père maternel "Djedou Saleh" et ma grand-mère maternelle "Mani Tamma", j'imagine que serait votre joie aujourd'hui, que Dieu vous accueille dans son vaste paradis.

À tous mes cousins et mes cousines, mes oncles et tantes.

À mes camarades et amis que j'estime, particulièrement, mon meilleur ami et mon binôme Manou, à tous les moments qu'on a partagés ensemble, ainsi que Brahim, Samir, Rayan, Walid, Hynd, Thafath et Souad, je vous souhaite énormément de réussite dans vos vies et que nos liens d'amitié se pérennisent à tout jamais.

Annis

Dédicaces

Je tiens à dédier ce travail à ma mère, cette sacrée personne qu'elle est, qui s'est tant sacrifiée pour voir son fils réussir et atteindre ses objectifs,

À mon père, qui a longtemps été garant de mon développement et mon évolution sur les bases qui ont fait de moi l'homme que je suis aujourd'hui.

À ma chère sœur Yasmine, qui a été là depuis le début à suivre mon parcours de près,

À mes chers frères Ilyes et Akram, que je leur souhaite tout le succès.

À mes grands-parents, que Dieu vous protège.

À la mémoire de mon grand-père Said, que Dieu t'accueille dans son vaste paradis.

À mes cousins et cousines, tantes et oncles,

À mon binôme depuis la première année, mon frère et mon ami, Anis, qui a toujours été à côté de moi pour relever plusieurs défis et aujourd'hui dans la réalisation de ce projet.

À l'ensemble de mes camarades de la promotion du Génie Indus, spécialement, Brahim, Samir, Rayan, Walid et Hynd.

À tous les Almuni qui m'ont conseillé, guidé et donné de leur temps entre autre Mahyou et Rostane.

Manou

Remerciements

Nous adressons nos plus vifs remerciements au Dr. Iskander Zouaghi pour les précieux conseils qu'il nous a prodigués, le temps qu'il nous a accordé et son encadrement rigoureux et continu qui nous a permis de réaliser ce modeste travail.

Nous tenons aussi à remercier la présidente du jury, Mme. Bouchafaa, et l'examinateur, M. Arki, de nous avoir honoré de leur temps et de leur savoir-faire pour l'évaluation de ce projet de fin d'études.

Nous remercions profondément notre promotrice au sein de Schlumberger N.A.F, Mme. Hadia Ouafi du département Import-Export, de nous avoir aidé à développer cette problématique et cette solution et nous avoir guidé tout au long de notre stage à Schlumberger N.A.F, nous remercions également M. Haddadi et M. Laib pour leur précieuse assistance durant notre stage.

Nous tenons aussi à remercier l'ensemble de nos enseignants du département du Génie Industriel pour nous avoir fait part de leur savoir-faire et connaissances tout au long de notre formation.

Une pensée particulière à tous les étudiants du département, notamment le club IEC pour ces trois merveilleuses années remplies de bons souvenirs.

Pour finir, nous remercions, tous ceux qui ont participé, de près ou de loin, à la concrétisation de ce projet.

Table des matières

Liste des tableaux

Table des figures

Liste des Abréviations

ABM Agent Based Modeling

ATA Actual Time of Arrival

BI Business Intelligence

CCA Customs Clearance Agent

CFD Complete File Date

CNN Convolutional Neural Network

CRISP-DM Cross Industry Standard Process for Data Mining

CTGAN Conditional Tabular Generative Adversarial Networks

DES Discrete Events Simulation

DM Modélisation dimensionnelle

DSC Distribution Service Center

ERP Enterprise Resource Planning

ETA Estimated Time of Arrival

ETL Export, Transform, Load

FCL Full Container Load

FD Full Duties

GAN Generative Adversarial Networks

GBDT Gradient Boosted Decision Trees

GOLD Global Oilfield Logistics and Distribution

HTC Harmonized Tariff Code

I/E Import/Export

KPI Key Performance Indicator

LCL Less-Than-Container-Load

LCT Logistic Control Tower

MOT Mean Of Transport

NAF North Africa

PIB Produit intérieur brut

RNN Recurrent Neural Network

SC Supply Chain

SD System Dynamincs

SLB Schlumberger

SMA Systèmes multi agents

SVM Support Vector Machine

TVA Taxe sur la valeur ajoutée

TVAE Tabular Variational Autoencoder

Introduction Générale

Introduction Générale

La nature du marché des services pétroliers impose une dimension internationale à leur supply chain, les entreprises de ce secteur opèrent sur plusieurs territoires à travers le monde et nécessitent d'acheminer des équipements et des produits tout au long de leur supply chain globale.

L'importation est donc une activité importante de ces entreprises et sa bonne gestion permet de minimiser les lead times et les coûts. Néanmoins, maitriser ce processus reste complexe due aux différents acteurs tel que la douane, les transitaires et les entreprises de transport, ainsi que les différentes étapes de ce processus, chaque étape engendre un lead time et un certain coût associé.

Schlumberger, étant une entreprise leader des services pétroliers et la première dans son secteur, opère sur plus de 120 pays et effectue quotidiennement des opérations d'importation de ses centres de distributions à travers le monde vers ses bases opérationnelles dans plusieurs territoires. La GeoUnit Schlumberger NAF située à Alger, importe fréquemment des quatre coins du monde vers ses bases opérationnelles à Hassi Messaoud, cette activité est une source importante des lead times et des coûts pour l'entreprise, ce qui a poussé le département d'import-export à chercher un moyen pour mieux planifier ces importations afin de prendre les bonnes décisions pour chaque expédition.

Cependant, il n'existe aucun outil qui permet à l'entreprise de bien planifier ces importations, plus précisément, l'estimation des lead times pour différentes situations d'importation permettra aux spécialistes import-export de l'entreprise de choisir le meilleur scénario d'importation en termes de plusieurs paramètres tel que le moyen de transport à utiliser, le port d'entrée à choisir, le transitaire à employer et d'autres paramètres, de plus l'estimation des lead times relatifs à l'importation permettra à l'entreprise d'estimer une partie importante des coûts de sa supply chain.

Avec le faible volume des données, la forte volatilité des lead times et leur dépendance de plusieurs facteurs, comment pourrions-nous développer un moyen efficace qui aidera à mieux estimer ces lead-times sur plusieurs scénarios ?

Notre projet a pour but de résoudre cette problématique en proposant un outil d'estimation de quatre lead times différents : le lead time de préparation du dossier de déclaration de l'expédition par une Business Line donnée, le lead time que prend un transitaire pour déclarer une expédition au niveau de la douane, le lead time que prend la douane pour traiter un dossier relatif à une expédition et enfin le lead time nécessaire pour la livraison d'une expédition à partir des différents entrepôts douaniers vers les bases opérationnelles de Schlumberger NAF. Par la suite, nous proposons aussi un simple modèle pour estimer les coûts de stockage dans les entrepôts douaniers ainsi que le coût de transport en se basant sur ces lead times estimées.

La nature stochastique des lead times nous a poussé à choisir l'apprentissage machine

pour la modélisation et la référence des projets de data mining CRISP-DM (*Cross Industry Standard Process for Data Mining*) qui nous servira comme méthodologie de travail, de plus, nous présenterons deux techniques de génération de données pour remédier au problème du faible volume de données et nous évaluerons nos modèles pour chaque situation.

Ce travail est composé de quatre chapitres :

I Le premier chapitre s'intéresse à la présentation des concepts clés relatifs aux supply chains internationales des services pétroliers et plus particulièrement les concepts et travaux liés à l'importation, ses acteurs et ses lead times. Nous présenterons aussi à la fin de ce chapitre des travaux dans la littérature relatifs à l'estimation des lead times en utilisant plusieurs techniques déterministes et stochastiques dans des contextes industriels différent (fabrication, livraison, etc).

II Le deuxième chapitre présentera les concepts clés liés à la Business Intelligence qui nous permettrons de développer des tableaux de bords pour mieux analyser le processus d'importation de l'entreprise ainsi que la Data Science, notamment les étapes de la méthodologie CRISP-DM, les différents modèles de l'apprentissage machine utilisés dans ce projet ainsi que les différentes techniques de génération de données.

III Dans le troisième chapitre, on présentera Schlumberger, plus particulièrement la Geo Unit "*North Africa*" ainsi que sa supply chain, par la suite, on définira le processus d'importation de l'entreprise avec ses différentes étapes.

IV Enfin, la solution sera implémentée dans le quatrième et dernier chapitre, où on détaillera les étapes de la mise en place de notre solution selon les différentes phases de la méthodologie CRISP-DM : la compréhension des métiers, la compréhension et la préparation des données ainsi que leur génération par deux techniques différentes, la modélisation des lead times et l'évaluation des modèles entraînés et enfin le déploiement de la solution.

Le résultat de notre travail est un Chat-bot interactif qui permet aux spécialistes de l'import-export de Schlumberger NAF d'estimer les coûts et les lead times de chaque expédition avant son arrivée en Algérie ainsi qu'un tableau de bord sur PowerBI pour analyser les performances de l'importation et un système de calcul des pénalités pour les transitaires Pour conclure, nous donnerons des améliorations potentielles pour notre solution et qui pourrait faire l'objet d'autres problématiques à traiter dans le futur.

Chapitre 1 : Lead Time des opérations d'importation dans les supply chains du secteur des services pétroliers

Chapitre 1

Lead Time des opérations d'importation dans les supply chains du secteur des services pétroliers

Afin de bien cerner la problématique traitée par ce travail, nous allons présenter la supply chain internationale dans le secteur des services pétroliers et son fonctionnement d'une manière générale. Nous allons par la suite détailler les opérations d'importation dans les supply chains internationales et introduire quelques concepts clés liés à cette dernière. Par la suite, nous explorons la notion de lead times relatifs à l'importation dans le secteur des services pétroliers.

1.1 La Supply Chain Internationale

La globalisation de l'économie et la facilitation des échanges commerciaux à travers le monde ont poussé plusieurs entreprises à se mondialiser, ouvrant par cela des marchés et des opportunités uniques, mais aussi des challenges et de nouvelles contraintes à prendre en considération. Quand une entreprise commence à opérer sur plus d'un seul territoire, sa supply chain devra s'adapter, il faudra mettre en place un nouveau réseau de distribution, de nouveaux modes de transport et renouveler sa stratégie afin de prendre en considération les différents challenges associés à une globalisation.

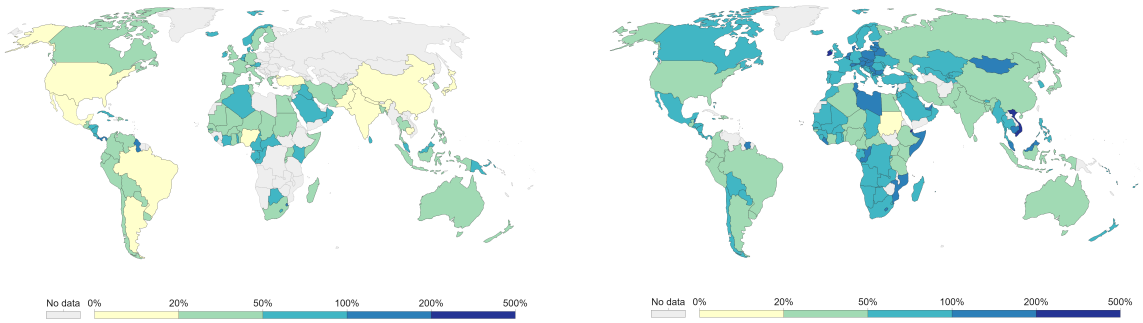
1.1.1 Management de la supply chain internationale

Le management de la supply chain internationale (*Global supply chain management*) a pour but de gérer les supply chains qui s'étalent sur plusieurs pays et territoires tout en minimisant les coûts et les lead times. Ce management prend en considération les régulations et procédures d'organismes gouvernementaux en plus des différents acteurs de la supply chain classique.

La stratégie de la supply chain adoptée par les entreprises multinationales se base principalement sur une intégration renforcée des fournisseurs et des clients, ainsi que sur une coordination accrue des multiples processus de création de valeur. De plus, cette stratégie nécessite une gestion des flux d'informations, matériels et financiers sur un niveau international, ce qui implique des décisions organisationnelles et financières importantes (Cohen & Huchzermeier, 1999). L'acheminement des produits et les différentes ressources

d'une entreprise multinationale entre plusieurs frontières nécessite une planification préalable qui devra prendre en considération les régulations et les procédures douanières de chaque pays. Une supply chain internationale qui n'intègre pas efficacement les services douaniers dans sa structure peut subir des coûts et des délais (lead times) supplémentaires qui influencent sa réactivité.

La gestion de supply chain internationale permet de traiter plusieurs problématiques qu'on développera par la suite, notamment le problème d'estimation et d'optimisation des lead times, qui présente un problème important dans les supply chains globales.



(a) En 1970 (b) En 2020

FIGURE 1.1 – Évolution des échanges commerciaux internationaux autant que pourcentage du PIB de 1970 à 2020 (World Bank, 2020)

Maintenant que nous avons présenté la supply chain internationale ainsi que sa gestion, nous allons nous focaliser dans la suite de ce chapitre sur la supply chain des services pétroliers au niveau international ainsi que ses composantes et ses acteurs.

1.1.2 Supply chain des services pétroliers

La supply chain du pétrole et du gaz peut être décomposée en trois parties, à savoir la partie « upstream » qui comporte la production et l'extraction des hydrocarbures, la partie « midstream » qui intègre le transport et la distribution vers les points de ventes et enfin la partie « downstream » qui concerne les points de ventes et les clients finaux, ainsi que les raffineries. Le secteur des services pétroliers opère principalement sur la partie « upstream » de la supply chain du pétrole et du gaz en offrant des services de forage, de production et de maintenance des puits de pétrole.

Les supply chains des entreprises de services pétroliers sont globales par nature, elles fournissent des équipements et des produits aux entreprises pétrolières et pour cela, leur configuration suit essentiellement un modèle "Hub and Spoke" ou l'entreprise dispose d'entrepôts ou de hub pour effectuer du cross-docking : les commandes arrivent à ces hubs, elles sont ensuite consolidées en plusieurs commandes est envoyé vers leurs destinations finales (Vanajakumari et al., 2022). Cette configuration permet de réduire les coûts de transport et de stockage en plaçant ces hubs d'une manière stratégique dans des zones d'échange libre où l'intervention des douanes n'est pas requise.

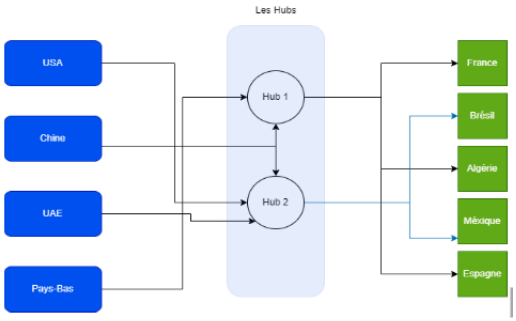


FIGURE 1.2 – Exemple d'une configuration "Hub & Spoke" à deux hubs

La nature globale de ces supply chain implique la nécessité d'intégrer les procédures douanières dans la planification de la supply chain, plus particulièrement, les équipements qui circulent tout au long de ce type de supply chain peuvent faire l'objet de régulation particulière. Par exemple, certaines entreprises de services pétroliers importent de la dynamite et des explosifs pour le forage, ce type de produit prend plus de temps à être dédouané et son transport est délicat vers les sites et les installations opérationnelles peut prendre plus de temps. Les lead times dans ce secteur sont donc importants et leur réduction permet de diminuer les coûts de manière significative.

1.1.3 Défis des supply chains globales des services pétroliers

Une supply chain globale des services pétroliers, fait face à plusieurs défis, les trois principaux défis sont : Le manque de ressources humaines à cause des différentes régulations sur le travail dans chaque pays et la démographie différente entre les pays, l'indisponibilité des équipements nécessaires pour subvenir au besoin d'une supply chain globale et l'effet boule de neige des goulots à travers la supply chain qui peut amplifier les lead times et causer des retards sur toute la chaîne (McKinsey, 2022).

Dans ce secteur, les types d'expéditions effectuées varient considérablement : tuyaux, vannes, grues, produits chimiques, explosifs, matériaux de construction, et les appareils de forage, de plus ces équipements et produits sont déplacés fréquemment entre différents pays, ce qui oblige les entreprises de services pétroliers à faire face au défi de prévoir et réduire les lead times de la distribution à cause des procédures douanières parfois compliquées et bureaucratiques ainsi que l'emplacement distant des sites et bases opérationnels (Sahara, Offshore, cercle arctique, etc).

D'autres défis sont aussi présents tels que la volatilité des taux de changes entre les monnaies, une faible variation du taux de changes peut s'amplifier et causer des coûts importants pour l'entreprise, un autre défi est le respect de qualité et de norme entre les différents fournisseurs, une supply chain globale devra intégrer des fournisseurs de plusieurs régions à travers le monde tout en s'assurant de la qualité des produits et ressources fournis. Certaines actions peuvent être mises en place afin de résoudre ces défis ou du moins minimiser leur impact sur les performances de la supply chain, Les boites de conseil KPMG (2021) et McKinsey (2022) proposent les solutions suivantes :

- Établir une collaboration plus étroite et renforcer l'intégration entre les différentes parties prenante de la supply chain internationale pour minimiser les risques dans ce secteur.
- Automatiser différents éléments de la supply chain comme l'entreposage, le déchargement, etc.

- Employer les données et les nouvelles technologies pour améliorer la visibilité à travers la chaîne et l'employer dans la planification pour prévoir et réduire les coûts et les lead times.

1.2 Les opérations d'importation dans la supply chain internationale

Dans cette partie, nous allons présenter les opérations et les concepts clés liés au processus d'importation et d'exportation, ses parties prenantes, les différents régimes douaniers ainsi que les principaux modes de transport employés. L'importation est l'acquisition des marchandises d'un pays étranger à des fins commerciales, elles peuvent être reçues par des particuliers, des entreprises ou des gouvernements.

1.2.1 Parties prenantes au processus d'importation et exportation

Durant une importation, il existe un nombre important d'acteurs qui prennent part dans ce processus, parmi ces parties prenantes on retrouve les sociétés de transport de marchandise, la douane et les transitaires :

- **Société de transport de marchandises** : Une société de transport de marchandises (Freight forwarders) est une entreprise qui fournit des services de transport avec toutes ses formalités (Lowe, 2002, p. 101). Ces entreprises possèdent des flottes importantes et permettent aux entreprises d'éviter de créer leur propre flotte de transport, une opération qui est très coûteuse. La société de transport peut être une compagnie qui opère sa propre flotte comme CMA-CGM ou encore une entreprise qui travaille en partenariat avec des compagnies maritimes, aériennes ou terrestres comme DHL.
- **La douane** : La douane (*Customs*) est une autorité gouvernementale qui est responsable de la mise en place et l'établissement de lois et de législations concernant l'importation et l'exportation dans un pays ainsi que la récolte des taxes et frais reliés à ces opérations (World Customs Organization, 2018, p. 9). En Algérie, les services douaniers sont régis par la direction générale de la douane qui est sous la tutelle du ministère des Finances.
- **Les Transitaires** : Un transitaire (*Customs Clearance Agent - CCA*) ou encore commissionnaire en douane, est une entreprise tierce ou un particulier qui fournit des services d'assistance aux autres entreprises durant le processus d'importation et exportation. Il est autorisé par l'autorité douanière d'un pays a exercé sur leur territoire et effectue plusieurs tâches administratives relatives aux formalités douanières comme la déclaration de marchandises importées. Il représente l'entreprise importatrice au niveau de la douane et lui permet de gagner du temps (World Customs Organization, 2018, p. 10). En Algérie, on retrouve plusieurs fournisseurs tels que la multinationale ARAMEX, ou d'autres entreprises nationales comme FENNEC, MEMORIAL TRANSIT, TRANSIT ACTION, etc. Le site web de la direction générale de la douane algérienne contient toutes les informations nécessaires sur les formalités, les obligations et les conditions nécessaires pour être habilité en tant que transitaire/commissionnaire en douane.

Dans une supply chain internationale, il est nécessaire d'intégrer de près ces acteurs afin de minimiser les coûts et éviter des lead times élevés. En plus de ces acteurs, il existe

aussi deux régimes d'importation différents qu'on détaillera par la suite.

1.2.2 Le processus d'importation et ses régimes

Le processus d'importation dépend du pays concerné, chaque pays comporte une certaine procédure à suivre pour importer de la marchandise, nous allons présenter brièvement, les étapes d'importation d'une marchandise en Algérie et ensuite nous explorons les différents régimes douaniers.

Le processus d'importation en Algérie depuis un autre pays se divise en 7 étapes principales (Direction Générale des Douanes, 2022) :

FIGURE 1.3 – Les différentes étapes du Processus d'importation

1. **Expression du besoin :** Tout d'abord, l'entreprise va exprimer son besoin en un certain ou plusieurs produits qu'elle souhaite acheter. Par la suite, elle choisit un fournisseur de sa liste de fournisseurs préalable à travers le monde, ou bien elle intègre un nouveau fournisseur. Un ensemble d'échanges et de négociations s'effectuent entre l'acheteur et le fournisseur.
2. **Établissement du contrat & incotermes :** Après un accord mutuel des deux parties, un contrat est établi entre l'acheteur et le vendeur, ce contrat stipule les mesures et les conditions générales de vente et d'achat. Dans ces contrats, on retrouve des incotermes : des codes normalisés qui précisent la responsabilité de l'acheteur et du vendeur par rapport à plusieurs aspects : le chargement, le transport, le type de transport, etc. Un incoterme qu'on retrouve souvent dans notre cas d'application (le secteurs des services pétroliers en Algérie) est le DAT (*Delivered At Terminal*). Ce dernier stipule que le vendeur prend en charge l'emballage, le chargement et la livraison à travers le moyen de transport principal jusqu'au terminal (un port ou un aéroport par exemple), cependant l'acheteur est responsable du dédouanement et de la livraison vers la destination finale. Dans ce qui suit, nous faisons l'hypothèse d'un contrat avec DAT comme incoterme.
3. **Chargement et embarquement :** Par la suite, le vendeur charge la marchandise et l'embarque dans le moyen de transport principal (bateau pour le fret maritime, avion/avion-cargo pour le fret aérien, camions pour le fret terrestre). Le vendeur reste responsable jusqu'à l'arrivée au terminal (port, aéroport ou autres) (Sous l'incoterm DAT).
4. **Arrivée de la marchandise et déchargement :** Après l'arrivée de la marchandise au port, dans le cas du fret maritime, le capitaine du navire devra effectuer un ensemble de formalité auprès de la douane, dans le cas du fret aérien, il suffit de déposer une déclaration de cargaison, tandis que dans le cas de fret terrestre, les camions ont le droit de circuler sans permis jusqu'au bureau de douane le plus proche du poste frontalier. Par la suite, le déchargement peut être effectué sous la présence d'agents de douane, la marchandise est transférée ensuite vers des entrepôts ou des magasins en attente du dédouanement.
5. **Procédures de dédouanement :** Une fois la marchandise arrivée au territoire de l'acheteur, elle est dans l'obligation de passer par la douane, l'acheteur fait appel

à un commissaire en douane ou un transitaire (personne physique ou morale) pour entreprendre les différentes procédures douanières :

(a) D'abord, le transitaire devra recevoir des documents sur la marchandise de la part de l'acheteur

(b) Une fois ces documents reçus, le transitaire ou le commissionnaire en douane déclare la marchandise en détail au niveau de la douane, la douane algérienne lui donne un délai de 21 jours pour déclarer la marchandise, en attendant, elle demeure dans des magasins sous la tutelle de la douane.

(c) La douane reçoit la déclaration en détail, un agent, dit agent de recevabilité en douane, vérifie les documents fournis ainsi que l'adéquation du régime choisi, en cas de problème, il recontacte le transitaire (ou la partie qui a effectué la déclaration) pour finaliser le dossier.

(d) Une fois que le document est complet et contrôlé par la douane, la déclaration est enregistrée et un numéro d'enregistrement unique est transmis au transitaire. Par la suite, un circuit est lancé (un circuit vert, orange ou rouge) selon la décision prise par les services douaniers.

(e) Pour finir, l'acheteur doit payer les frais et taxes douaniers soit directement, soit à travers le transitaire, le paiement peut être effectué en espèce ou autre moyen de paiement. Pour le secteur des services pétroliers, certains avantages fiscaux ont été mis en place par le gouvernement Algérien, ces avantages sont détaillés à la fin de cette partie.

6. **Le bon à enlever :** Après le paiement des frais et taxes douaniers, le dédouanement est officiellement terminé, la douane envoie un bon à enlever au transitaire qui lui permet de faire sortir la marchandise légalement.

7. **Livraison à la destination finale :** Par la suite, la livraison à la destination finale se fait selon le choix de l'acheteur : il peut opter pour les services de transport du transitaire, opter pour des services logistiques d'une autre entreprise ou bien s'occuper lui-même du transport de la marchandise.

Cependant, avant d'importer une marchandise, il est impératif de sélectionner le régime douanier adéquat. Il existe deux principaux régimes d'importation en Algérie, en l'occurrence le régime temporaire et le régime permanent.

Importation temporaire (admission temporaire) : La douane algérienne permet d'importer de la marchandise sur le territoire douanier et le réexporte après un délai déterminé. Il existe deux types d'importation temporaires :

- **Importation temporaire avec réexportation en état :** ce type d'admission temporaire permet d'importer des équipements, de les utiliser pendant une certaine durée et de les exporter dans leur état initial. Parmi ces équipements, on retrouve les palettes, les conteneurs, les véhicules routiers commerciaux, les équipements de production, etc.
- **Importation temporaire avec perfectionnement :** ce régime est employé dans le cas d'importation temporaire d'un équipement et de sa transformation avant de le réimporter. Cette transformation peut se présenter sous forme d'une maintenance, réparation ou amélioration.

L'importation temporaire est principalement utilisée par les entreprises pour importer du matériel lourd ou des actifs importants qui peuvent être réutilisés par d'autres filiales dans d'autres pays par exemple.

Importation permanente : Le régime permanent s'applique à l'importation et l'exportation et permet d'importer ou exporter de la marchandise d'une manière permanente sans la réexporter ou réimporter, respectivement. C'est le régime standard que la majorité des entreprises envisage pour la marchandise à consommation ou même certains équipements.

Enfin, il est à souligner que le secteur pétrolier bénéficie d'avantages fiscaux pour certaines activités telles que la recherche, la prospection et l'exploitation des hydrocarbures selon l'article 58 de la loi n°86-14 du 19 Août 1986 (modifiée et complétée par la loi n°91- 21 du 4 Décembre 1991). La taxe de valeur ajoutée (TVA) est exemptée dans ce cas. Cette loi s'applique donc sur les entreprises d'exploitation pétrolière aussi ainsi que les entreprises de service pétrolier.

En plus de ces régimes douaniers, les moyens de transport utilisés sont une composante importante de la supply chain internationale, dans ce qui suit, nous présenteront les types de transports utilisés et leurs caractéristiques.

1.2.3 Les moyens de transports utilisés au niveau international

Lors des échanges internationaux, plusieurs modes de transport sont utilisés pour transporter les marchandises d'un endroit à l'autre. La majorité des marchandises sont expédiées par fret : maritime, terrestre ou aérien. Nous allons présenter brièvement chaque type de fret selon Sarder (2021) :

- **Le fret maritime :** Le fret maritime est une méthode de transport de grandes quantités de produits via des cargos. Les marchandises sont emballées dans des conteneurs et ces conteneurs sont chargés sur un navire, où ils seront acheminés vers leur pays de destination. Ce mode assure actuellement le transport de 90% des marchandises non-vrac dans le monde. Il existe deux types d'expéditions LCL et FCL (Figure 1.4).
  - **LCL :** L'abréviation LCL signifie Less-Than-Container-Load et fait référence à des chargements partiels dans un conteneur. Donc des chargements de plusieurs expéditeurs seront envoyés dans un même conteneur consolidé.
  - **FCL :** Full Container Load, ou conteneur complet : les marchandises d'un client voyagent dans un conteneur rempli et scellé.

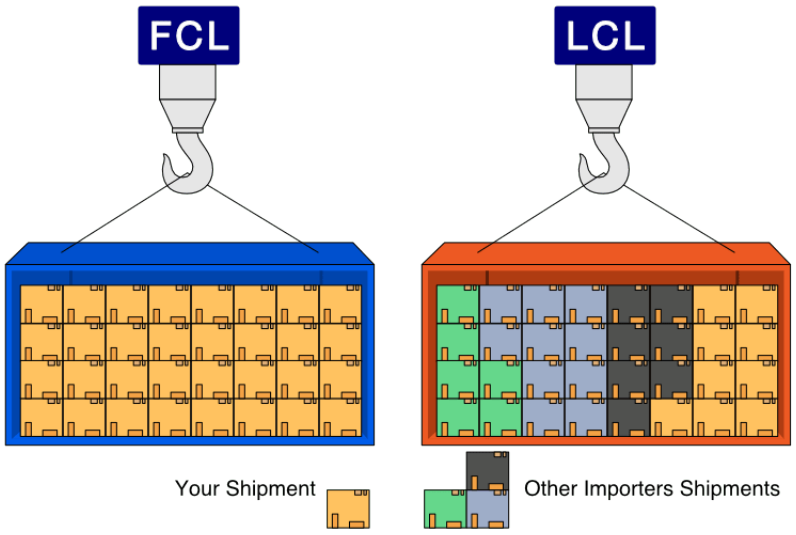


FIGURE 1.4 – Le régime FCL et LCL pour le transport maritime (Shipper, 2021)

- **Le fret aérien :** Ce mode de transport particulier est très coûteux par rapport aux autres modes. Cependant, le fret aérien est rapide et fiable. Le fret aérien est expédié soit dans l'espace supplémentaire des avions de ligne qui n'est pas occupé par les passagers et leurs bagages, soit par l'intermédiaire de sociétés de fret aérien spécialisées, comme UPS ou FedEx.
- **Le fret routier :** Ce mode de transport a connu la plus forte croissance au cours des 50 dernières années en raison de la libéralisation du commerce. L'expédition par camion a un coût relativement faible, car le secteur est très concurrentiel. Il est adapté pour le transport dans un même pays ou entre deux pays ayant des frontières terrestres.
- **Le transport intermodal et multimodal :** Les deux types de transport utilisent une combinaison de mode de transport afin de livrer les marchandises d'un point A vers un point B. La différence entre les deux types, c'est que dans le transport multimodal on trouve un seul contrat et l'entreprise de transport est la seule responsable des marchandises. Cependant, le transport intermodal, il existe un contrat distinct pour chaque étape du voyage. Cela signifie qu'il y a plus d'une entité responsable de la bonne livraison de la marchandise.

Après avoir découvert les différents moyens de transport ainsi que les régimes douaniers utilisés lors des échanges internationaux, nous allons à présent mettre en évidence l'importance des lead times dans les opérations d'importation.

1.2.4 Les lead times dans les opérations d'importation

Afin de rester compétitives, les entreprises adoptent diverses stratégies, telle que l'externalisation à l'étranger, cette stratégie expose les entreprises à une incertitude dans le transit (dédouanement) qui va impacter le lead time de la livraison (Colicchia et al., 2010). D'après l'étude réalisée par Baig et al. (2022), qui consiste à évaluer les capacités de résilience de la Supply Chain contre les vulnérabilités les plus importantes et les plus courantes dans le contexte de l'industrie pétrolière du Pakistan, ils ont trouvé que la première vulnérabilité est liée à l'offre et la demande (Demand and Supply).

Les lead times sont devenus de plus en plus importants, c'est pour cela que le terme résilience de la Supply Chain prend de l'ampleur. La résilience de la Supply Chain représente sa capacité d'adaptation à se préparer à des événements inattendus, à réagir à des perturbations et à s'en remettre en maintenant la continuité des opérations au niveau souhaité de connectivité et de contrôle de la structure et de la fonction (Ponomarov & Holcomb, 2009).

Les lead times constituent un élément encore plus important dans la supply chain des services pétroliers à cause de leur complexité et la nécessité d'être réactif à la demande croissante en hydrocarbures. De plus, la supply chain des services pétroliers est généralement globale, ce qui peut causer des lead times importants.

Après avoir détaillé les concepts clés relatifs à l'aspect supply chain de notre problématique, dans ce qui suit, nous allons présenter la Business Intelligence ainsi que la Data Science qui nous serviront de cadre de référence pour notre travail, et cela, à travers la méthodologie CRISP-DM.

Chapitre 2 : Déploiement de la Business Intelligence et de la Data Science dans la détermination des lead times dans les opérations d'importation

Chapitre 2

Déploiement de la Business Intelligence et de la Data Science dans la détermination des lead times dans les opérations d'importation

À travers ce chapitre, nous allons découvrir la Business Intelligence et la data science. Ces deux domaines sont complémentaires et essentiels dans l'estimation des lead times, la Business Intelligence se focalise sur le passé et le présent et permet de donner une analyse descriptive tant dis que la Data Science se focalise sur le futur et permet de réaliser une analyse prédictive.

2.1 Business Intelligence

La Business Intelligence (BI) est une combinaison de processus, de politiques, de culture et de technologies pour la collecte, la manipulation, le stockage et l'analyse de données provenant de sources internes et externes, afin de communiquer des informations, de créer des connaissances et d'éclairer la prise de décision. (Foley & Guillemette, 2010)

La BI permet de rendre compte des performances de l'entreprise, de découvrir de nouvelles opportunités commerciales et de prendre de meilleures décisions commerciales concernant les concurrents, les fournisseurs, les clients, les questions financières, les questions stratégiques, les produits et les services.

En ce qui suit, nous allons définir les différentes architectures de BI, la modélisation dimensionnelle et enfin les tableaux de bord et KPIs.

2.1.1 Architecture générale de la BI

En Business Intelligence il existe plusieurs architectures, les plus répandues sont celles de Inmon (2002) et Kimball et Ross (2013).

- **L'approche de Inmon :** c'est une approche Top - Down, ou il propose une architecture de données basée sur un Data Warehouse qui est définie par Inmon comme étant une collection de données orientée sujet, intégrée, variables dans le temps et non volatile en soutien au processus de prise de décision de la direction. Après avoir créé un Data Warehouse, plusieurs processus ETL seront déployés, ces

processus consistent à exporter, transformer et enfin charger les données afin de créer des Data Marts qui sont des versions réduites d'un Data Warehouse, ils sont orientés pour analyser un seul domaine d'activité. Nous pouvons avoir un Datamart pour le département Finance, un autre pour le département de Supply chain, etc. La figure ci-dessus représente l'architecture de Inmon :

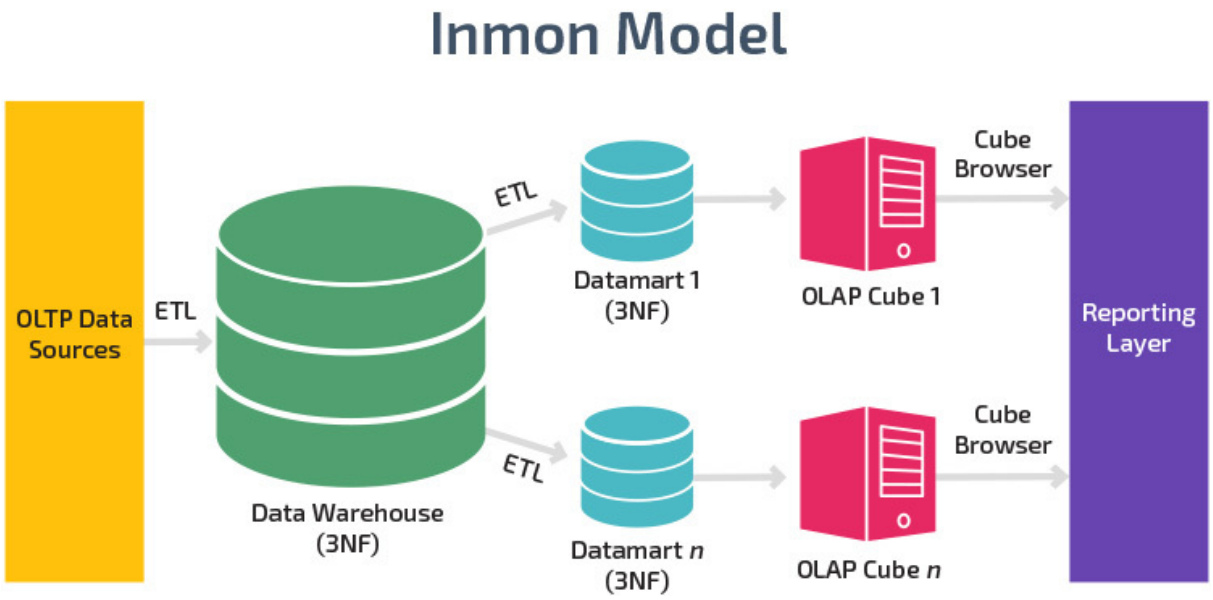


FIGURE 2.1 – Modélisation de Inmon (Panoply, s. d.)

- **L'approche de Kimball :** Kimball propose une approche qui débute par le design de plusieurs Data Marts adapté pour chaque processus Business ensuite le Data Warehouse sera créé par l'intégration des différents Datamarts. Comme montré dans la figure suivante.

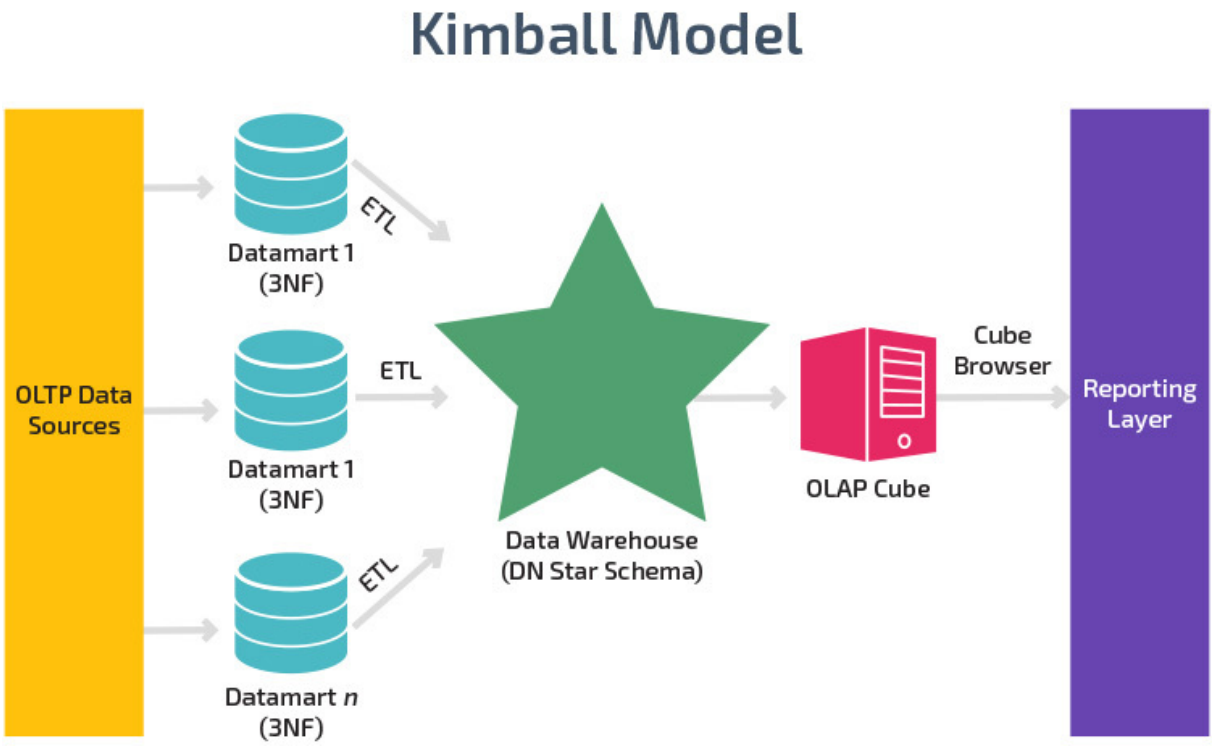


FIGURE 2.2 – Modélisation de Kimball (Panoply, s. d.)

Maintenant qu'on a découvert les architectures les plus utilisées, nous allons découvrir la modélisation dimensionnelle et ses étapes.

2.1.2 La modélisation dimensionnelle

La modélisation dimensionnelle (DM) est une technique de structure de données optimisée pour leur stockage dans un entrepôt de données. L'objectif de la modélisation dimensionnelle est d'optimiser la base de données pour une récupération plus rapide. Le concept de modélisation dimensionnelle a été développé par Ralph Kimball et consiste en des tables de faits et des dimensions.

En général, la modélisation dimensionnelle nécessite 5 étapes qui sont résumées dans le schéma ci-dessous :

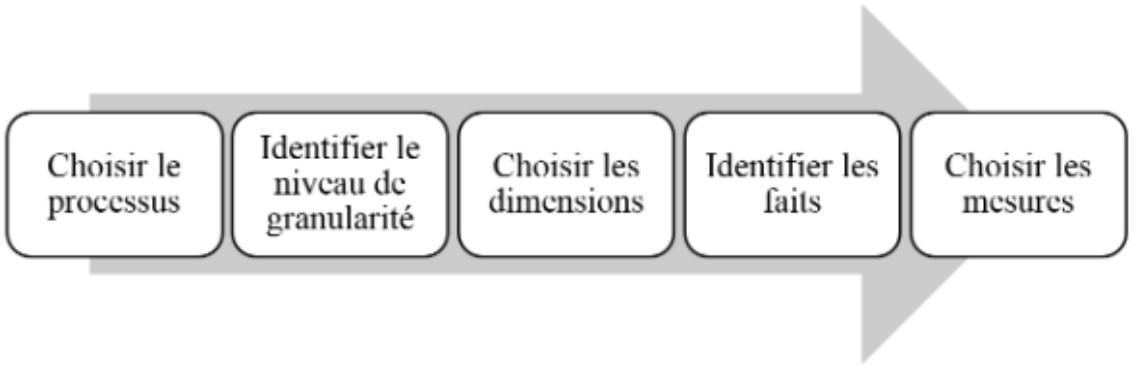


FIGURE 2.3 – les différentes étapes de la modélisation dimensionnelle

1. **Choix du processus :** la première étape de la conception consiste à décider du ou des processus métier à modéliser en combinant la compréhension des exigences métier et celle des données disponibles.
2. **Identification du niveau de granularité :** une fois le processus métier identifié, on doit prendre une décision importante concernant la granularité : quel niveau de détail des données doit être mis à disposition dans le modèle dimensionnel ? On peut directement choisir les données atomiques qui offrent une flexibilité analytique maximale, car on pourra faire tout type d'analyse, cependant, manipuler les données atomiques va affecter le temps d'exécution des requêtes, mais aussi les infrastructures nécessaires pour stocker les données.
3. **Choix des dimensions :** une fois que le niveau de granularité est fixé, les dimensions vont apparaître systématiquement.
4. **Identification des faits :** la quatrième étape de la conception consiste à déterminer avec soin les faits qui apparaîtront dans la table des faits. Une fois encore, la déclaration du niveau de grain nous aide à ancrer notre réflexion : les faits doivent être en concordance avec le niveau de granularité choisi.

Après avoir détaillé la modélisation dimensionnelle qui est nécessaire pour la création des tableaux de bord, nous allons maintenant détailler les tableaux de bord et KPIs.

2.1.3 Tableaux de bord et KPIs

Le tableau de bord est un instrument de pilotage d'entreprise. Il constitue un affichage visuel des informations les plus importantes et nécessaires pour réaliser des objectifs.

Ces informations sont consolidées et organisées sur un seul écran afin que les données soient visibles et puissent être surveillées instantanément. La figure suivante représente un exemple de tableau de bord :

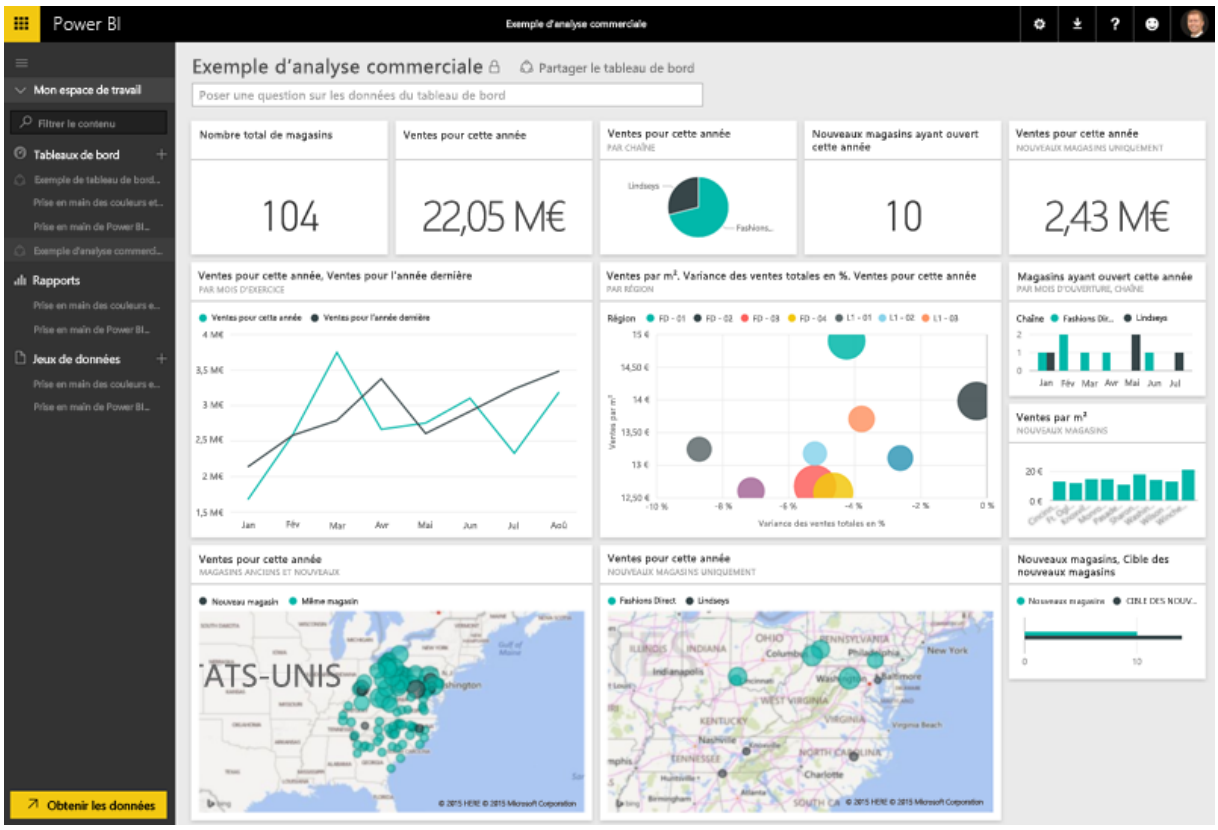


FIGURE 2.4 – Exemple d'un tableau de bord (Chaillou & Dugué, 2021)

Il existe quatre niveaux d'utilisation des tableaux de bord, le niveau le plus haut est le niveau stratégique, il est utilisé par les cadres supérieurs afin de suivre la stratégie à long terme de l'entreprise à l'aide des facteurs critiques de succès. Un niveau plus bas est le niveau analytique, il contient une grande quantité de données créées et utilisées par des analystes pour fournir un soutien aux dirigeants. Ensuite, nous avons le niveau tactique dont le but est l'analyse et le suivi des processus menés par les cadres intermédiaires. Enfin, nous trouvons les tableaux de bord opérationnels, ces tableaux sont utilisés pour la gestion des opérations qui ont un horizon temporel plus court qui sont gérés par des profils juniors de management (Pine, 2021).

Les tableaux de bord englobent plusieurs indicateurs, et d'après Parmenter (2020) les organisations confondent toujours entre les indicateurs en considérant tous les indicateurs comme des indicateurs de performance, c'est pour cela qu'il a proposé quatre types d'indicateurs qui sont : Indicateur clé de résultat, indicateur de résultat, indicateur de performance et indicateur clé de performance.

- **Les indicateurs de résultat :** ce sont des indicateurs qui résument l'activité de plusieurs équipes et qui sont adaptés pour avoir une vue globale du travail des équipes. Exemple : des indicateurs financiers, car ce sont des indicateurs de résultat.
- **Les indicateurs clés de résultat :** donnent un résumé global sur la performance de l'organisation, ils englobent les résultats de plusieurs équipes, et ils sont nommés

"clé" parce qu'ils résument la situation, exemple : le retour sur le capital employé, profit net avant impôt.

— **Les indicateurs clés de performance :** Les indicateurs clés de performance (KPI) sont les indicateurs qui se concentrent sur les aspects de la performance organisationnelle qui sont les plus critiques pour le succès actuel et futur de l'organisation.

— **Les indicateurs de performance :** sont les indicateurs non financiers (sinon ils seraient des indicateurs de résultat) qui peuvent être rattachés à une équipe. La différence entre les indicateurs de performance et les KPI est que ces derniers sont considérés comme fondamentaux pour le bien-être de l'organisation. Les indicateurs de performance sont importants, mais ne sont pas cruciaux pour l'entreprise, ils aident les équipes à s'aligner sur la stratégie de leur organisation.

Il existe plusieurs outils disponibles pour créer des tableaux de bord, allant du logiciel Excel jusqu'à l'utilisation d'un langage de programmation pour générer un tableau de bord. Les entreprises optent généralement pour des logiciels adaptés à la Business Intelligence comme PowerBi et Tableau.

Le logiciel Tableau est créé par l'entreprise Tableau, qui ensuite a été racheté par Salesforce, il offre une solution complète pour un projet de Business Intelligence, il existe plusieurs versions de ce dernier : Tableau Online, Tableau Desktop, Tableau Public et Tableau Server.

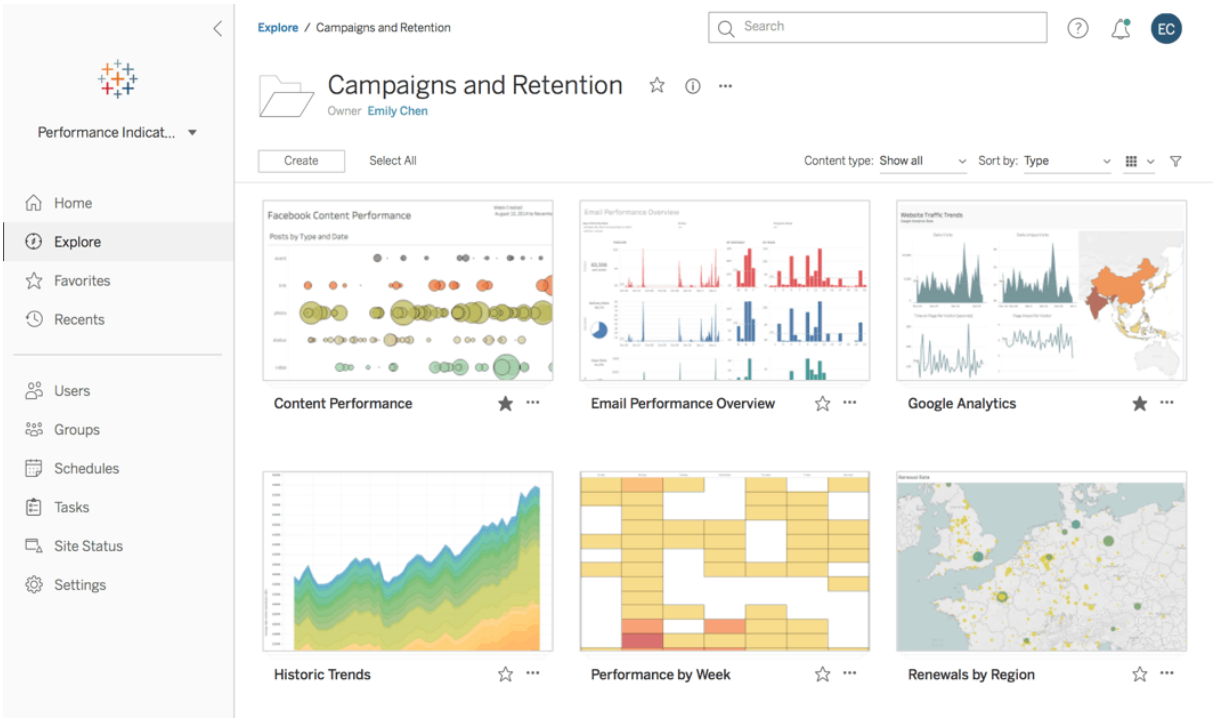


FIGURE 2.5 – Aperçu du logiciel Tableau (AnyLogic, 2022)

Power BI est une solution de Business Intelligence développée par Microsoft, il permet de créer des tableaux de bord et de les publier au sein d'une même organisation. Power BI englobe quatre grandes parties qui sont :

— L'intégration des données : il permet de connecter plusieurs sources de données en local ou en ligne via un seul clic

— Le processus ETL : le logiciel permet d'extraire, transformer puis charger les données afin de les utiliser dans les rapports.

— La modélisation en Data Warehouse : Power BI permet de modéliser les données sous un data warehouse afin de simplifier les analyses des données.
— Le reporting : il offre une interface simple qui permet d'insérer facilement des visualisations.

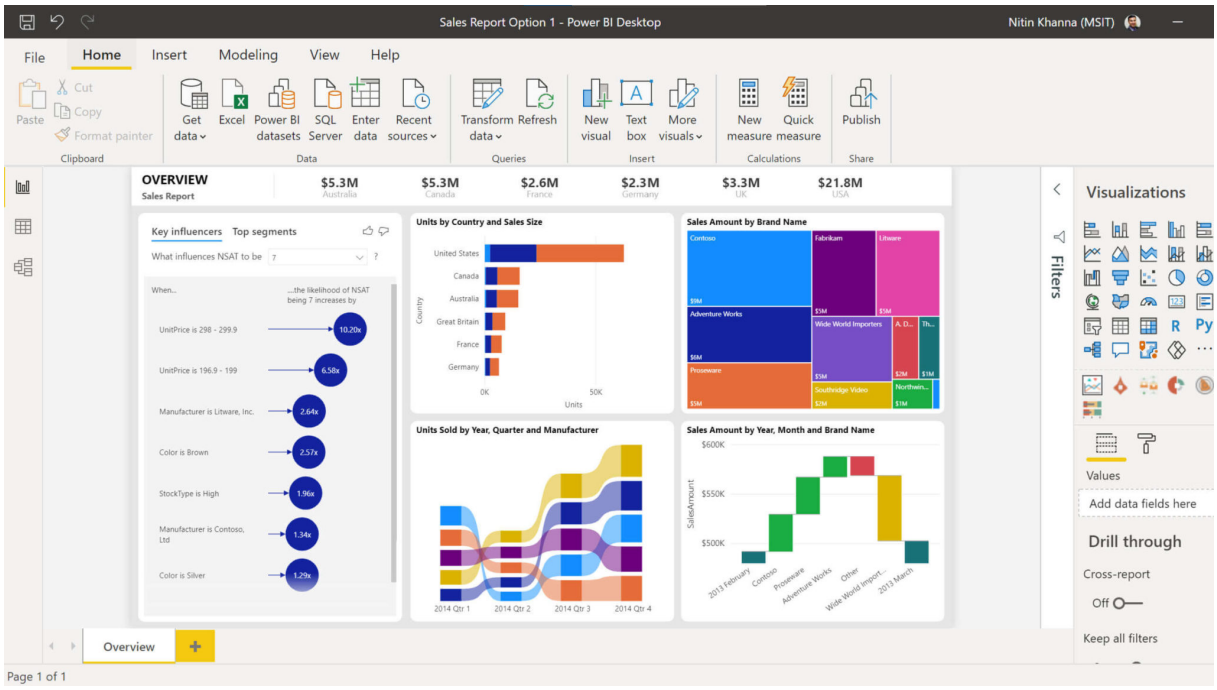


FIGURE 2.6 – Aperçu du logiciel Power BI (Jeremy, 2021)

Après avoir découvert les concepts liés à la Business Intelligence qui nous permettront d'analyser les données d'une manière descriptive, nous allons découvrir le domaine de la Data Science.

2.2 Data Science

La data science est une approche multidisciplinaire visant à obtenir des informations exploitables à partir des données. Elle englobe la préparation des données pour leur analyse et de leur traitement, le développement de modèle à travers le Machine Learning, et le déploiement de ces modèles. La data science permet de décrire le futur, c'est donc un outil d'analyse prédictive, néanmoins, elle partage certains concepts et objectifs en commun avec la Business Intelligence présenté dans cette figure :

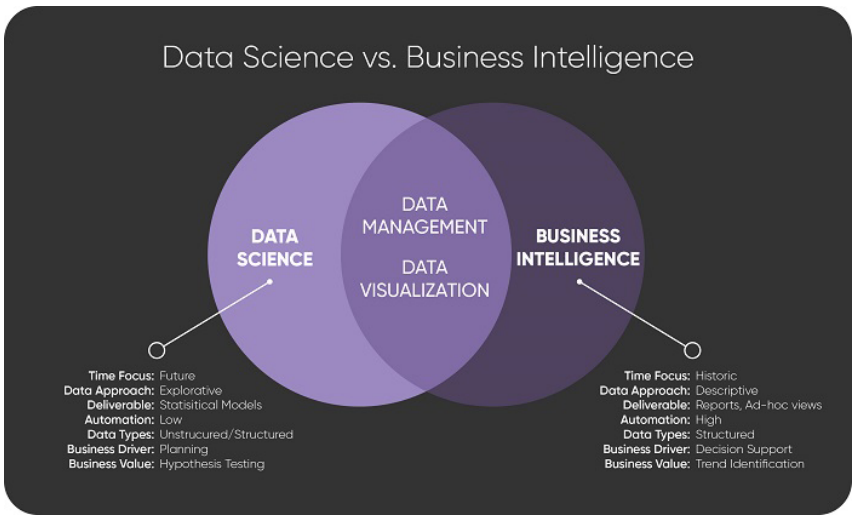


FIGURE 2.7 – Relation entre la Data Science et la Business Intelligence (KDnuggets, 2021)

Nous allons maintenant présenter une méthodologie qui servira comme cadre de travail pour notre solution connue sous le nom de CRISP-DM. Ensuite nous allons présenter la simulation, une méthode utilisée dans la modélisation des systèmes complexes, nous allons par la suite, définir le Machine Learning ainsi que son utilisation dans les problèmes de régression qu'on utilisera pour développer le modèle d'estimation des lead times. Enfin, nous définirons un ensemble de techniques utilisées pour la synthétisation des données.

2.2.1 La méthodologie CRISP-DM

CRISP-DM, acronyme de Cross Industry Standard Process for Data Mining, est une méthodologie développée par Chapman et al. en 2000. Vingt ans plus tard, la méthodologie est devenue la norme dans les projets de data mining et de data science. (Martínez-Plumed et al., 2021) Cette méthodologie englobe six grandes phases :

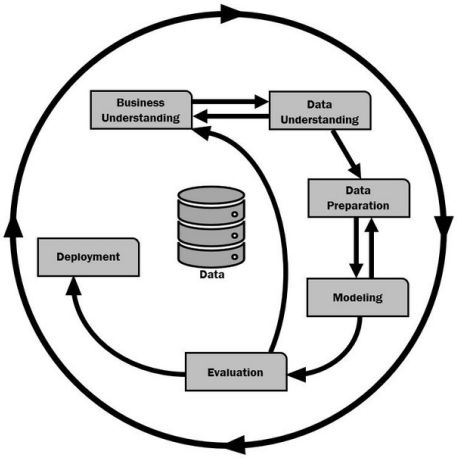


FIGURE 2.8 – Les étapes du CRISP-DM (Nima, 2019)

1. **Compréhension des métiers :** la première phase a pour but de définir l'objectif du projet d'un point de vue business, ensuite formuler une problématique d'exploration de données ainsi qu'un plan préliminaire pour atteindre ces objectifs.

2. **Compréhension des données :** parmi les tâches essentielles de cette phase, la collecte des données à partir des sources de données, leur exploration et leur description ainsi que la vérification de la qualité des données.
3. **Préparation des données :** cette phase consiste à préparer les données pour le modèle final. La préparation des données englobe plusieurs étapes comme le nettoyage des données, la construction de nouveaux attributs ainsi que la transformation des données.
4. **Modélisation :** cette phase à pour but de tester plusieurs modèles ainsi que la recherche des paramètres optimaux de chaque modèle, cette phase est fortement liée avec la phase de préparation des données.
5. **Évaluation :** après avoir essayé plusieurs modèles, il est nécessaire de les évaluer d'un point de vue business et non seulement d'une perspective d'analyse de données.
6. **Déploiement :** la création du modèle ne représente pas la fin du projet, car il faut le déployer et faciliter son exploitation par les utilisateurs.

Maintenant qu'on a défini les différentes étapes du CRISP-DM, nous allons maintenant découvrir la simulation et ses approches.

2.2.2 La simulation

La simulation est un outil informatique et mathématique puissant pour modéliser des systèmes complexes et dynamiques, elle permet de répliquer ces systèmes et suivre leur évolution à travers le temps et offre un certain degré de flexibilité pour modéliser la nature stochastique de ses propriétés. De plus, plusieurs outils et logiciels de simulations offrent la possibilité de visualiser la simulation à travers des animations.

Elle peut être employée pour trouver la configuration optimale d'un système complexe telle que la supply chain, de synthétiser des données et d'analyser les performances courantes d'un système, et cela, à plusieurs niveaux d'agrégation : niveau microscopique, niveau mésoscopique et le niveau macroscopique.

Il existe plusieurs approches de simulation selon le niveau de détails souhaité, parmis ces approches, on retrouve : La simulation à événements discrets (Discrete Events Simulation - DES) , la simulation à travers la dynamique des systèmes (System Dynamics - SD), la simulation basée sur les agents (Agent Based Modeling - ABM).

— **La simulation à événements discrets :** cette approche est utilisée pour effectuer des simulations au niveau microscopique d'un système, elle permet d'obtenir des informations détaillées de chaque composante du système simulé. Dans le cas de la simulation d'une supply chain par cette approche, on peut obtenir des informations détaillées sur les lead times, le nombre de commandes à chaque échelon, les différentes quantités, etc.

— **Simulation basée sur la dynamique des systèmes :** Le concept de dynamique des systèmes a été développé par Jay W. Forrester, c'est une approche qui propose un ensemble d'outils pour modéliser des systèmes complexes et dynamiques avec la notion de boucle de feedback (feedback loop), des flux pour modéliser la circulation d'objets matériels ou immatériels, des stocks pour modéliser le remplissage d'une quantité et autres outils (Sterman, 2009). La simulation à travers cette approche est fréquemment utilisée pour étudier le comportement d'un système à long terme et sur un niveau macroscopique.

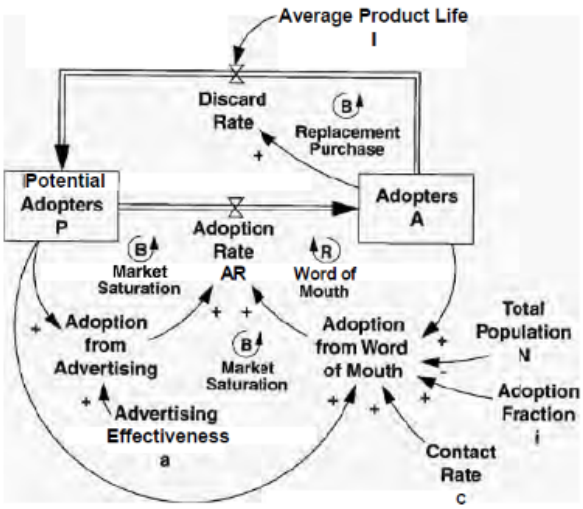


FIGURE 2.9 – Modélisation de l'adaptation de nouveau produit à la dynamique des systèmes (Sterman, 2009, p. 343)

— **Simulation basée sur les agents :** c'est une approche de simulation qui s'intéresse principalement aux agents comme étant les principaux éléments d'une simulation. Par exemple, dans une supply chain classique, on peut considérer le client comme un agent, le fournisseur comme un autre agent, l'entreprise comme un agent, etc. Chaque agent peut suivre une certaine logique de simulation et effectuer des actions qui modifieront son état ou ses paramètres. L'approche de modélisation par agent découle des systèmes multi agents (SMA), une branche d'intelligence artificielle qui permet de modéliser des systèmes complexes incorporant plusieurs acteurs et un environnement.

Après avoir découvert la simulation, nous passons maintenant au Machine Learning, une méthode très utilisée pour résoudre les problèmes de régression dont l'estimation des coûts et lead times.

2.2.3 Le Machine Learning et les problèmes de régression

L'apprentissage machine (Machine Learning), plus formellement, l'apprentissage statistique, est une branche de l'intelligence artificielle qui emploie les méthodes statistiques et des données afin de développer des modèles qui permettent de représenter un phénomène, un concept, un attribut, un résultat, etc. Elle essaye d'imiter la manière dont les humains apprennent à travers des algorithmes d'optimisation mathématique qui "apprennent" et s'améliorent graduellement à travers les données fournies.

Formellement, le but des algorithmes du machine learning est d'estimer une fonction f quelconque et de minimiser l'écart entre les données observées y_i et cette fonction est estimée. Il existe plusieurs types d'apprentissage, dans ce travail, nous allons présenter l'apprentissage supervisé et non supervisé et nous allons par la suite présenter un ensemble d'algorithmes d'apprentissage supervisé de régression.

Apprentissage supervisé

Dans l'apprentissage supervisé (Supervised learning), nous avons un échantillon de données d'entrée (variables exogènes) X ainsi que leur valeur correspondante pour la variable endogène y (données en sortie). Le but est donc d'estimer une fonction f qui permet de réduire l'erreur :

$$\hat{y} = \hat{f}(X) \tag{2.1}$$

$$y = \hat{y} + \epsilon = \hat{f}(X) + \epsilon \tag{2.2}$$

Dans ce cas, $\hat{f}$ est la fonction estimée et ϵ est l'écart entre les valeurs estimé avec cette fonction $\hat{y}$ et les valeurs réelles observées dans la variable endogène (variable de sortie) y. Cette erreur est composée d'une erreur réductible (qu'on peut réduire avec plus de calcul ou un meilleur algorithme d'apprentissage) et une erreur irréductible à cause de la nature stochastique du problème :

$$\epsilon = \epsilon_{\text{rd}} + \epsilon_{\text{irr}} \tag{2.3}$$

L'apprentissage supervisé permet de traiter deux types de problèmes :

- **Régression :** dans ce type de problème, le but est de développer un modèle de régression, la variable endogène (dépendante ou de sortie) y est donc un vecteur comportant des valeurs numériques. Le modèle appris permet d'obtenir un y_i pour une nouvelle entrée x. Un exemple est de développer un modèle de régression des prix des maisons par rapport à un ensemble de variables (surface, ville, sécurité, etc).
- **Classification :** dans ce type de problème, le but est de développer un modèle qui permet de classer chaque entrée de donnée x des données d'entrée X à une classe c_i d'un ensemble de classes Ω, dans ce cas, la variable y est catégorique (nominale ou ordinale) et prend des valeurs dans Ω, par exemple le développement d'un modèle qui peut prévoir si un client sera satisfait ou pas selon un ensemble de facteurs (variables exogènes), dans ce cas la classification est binaire, car $\Omega = \{\text{satisfait} = 1, \text{non} - \text{satisfait} = 0\}$. On peut aussi avoir des problèmes de classification multiple.

Parmi les algorithmes de régression, nous citons :

La régression linéaire : c'est l'algorithme le plus classique et le plus simple dans le machine learning. Le modèle est donnée par :

$$\hat{y} = \hat{w_1}x_1 + \hat{w_2}x_2 + ... + \hat{w_n}x_n + \hat{b}$$

Avec $\hat{y}$ étant la valeur estimée, w_j les poids (paramètres) du modèle qu'on veut estimer et x_j les valeurs de chaque variable j (dimension) pour une ligne des données disponibles, On peut reformuler avec une notation vectorielle plus compacte à travers le produit scalaire entre $\hat{w}$ (vecteur des poids qu'on souhaite estimer) ainsi que x :

$$\hat{y} = \hat{w} \cdot x + b$$

L'estimation des poids $\hat{w}$ peut s'effectuer à travers plusieurs méthodes tel que la méthode des moindres carrées ordinaires ou bien le maximum de vraisemblance. Les deux méthodes donnent la même formule exacte pour obtenir la meilleure droite, plan ou hyper-plan

(selon la dimension des données) qui minimise l'écart entre la valeur estimé $\hat{y}_i$ et la valeur théorique y_i. Cette formule est donnée par (sous forme matricielle) :

$$\hat{w} = (X^t X)^{-1} X^t y$$

L'annexe B contient plus de détails sur les conditions nécessaires de la régression linéaire.

Support Vector Regression : l'algorithme de support vector machine (SVM), développé en 1963, est un algorithme robuste et très utilisé en classification, il peut être aussi utilisé en tant que modèle de régression.

Le principe est de minimiser le module des coefficients $\|w\|_2$ (norme L_2) sous la contrainte que l'erreur absolue $|y - w_i x_i|$ est comprise dans une certaine marge d'erreur ϵ qu'on peut définir :

$$\begin{array}{ll} \min & \frac{1}{2}\|w\| \\ \text{s.c} & |y_i - w_i x_i| \leq \epsilon, i = 1, \ldots, n \end{array}$$

Ce programme mathématique (quadratique) peut être résolu à travers la méthode de Lagrange ou autres méthodes, et il permet d'obtenir deux hyperplans (droite en $\mathbb{R}^2$, plans en $\mathbb{R}^3$ et hyperplans en $\mathbb{R}^d, d \geq 3$), regroupant le maximum de données comme le montre la figure 2.10 :

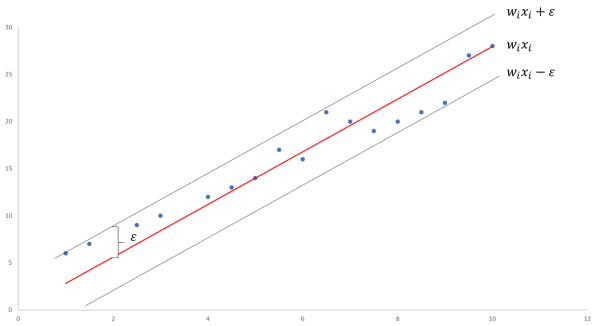


FIGURE 2.10 – La marge d'erreur et la droite (hyperplan) calculé par l'algorithme du SVM

Il est possible de réduire l'erreur d'avantage en introduisant des variables d'écarts ξ_i et un coefficient de tolérance C, le programme de minimisation quadratique devient par la suite :

$$\begin{array}{ll} \min & \frac{1}{2}\|w\| + C\sum_{i=0}^{n} |\xi_i| \\ \text{s.c} & |y_i - w_i x_i| \leq \epsilon + |\xi_i|, i = 1, \ldots, n \end{array}$$

Le modèle SVM permet aussi de modéliser des relations non linéaires en introduisant l'astuce du noyau (Kernel trick) développé en 1995. Le principe de cette méthode est de remplacer x_i par une fonction $\phi(x_i)$ non linéaire et d'utiliser le dual du programme d'optimisation présenté dans cette partie.

Decision trees : les arbres de décision (decision trees) sont des algorithmes de classification performants qu'on peut employer en régression aussi.

Le principe de cet algorithme est de construire un arbre binaire T, avec une certaine profondeur prédéfinie (à chaque niveau, nous avons deux coupes vu que l'arbre est binaire), chaque nœud k de l'arbre est divisé selon une dimension k et un seuil t_k. Le couple (k, t_k) est trouvé en minimisant la fonction objective suivante (connu sous le nom de CART : "Classification and Regression Trees") :

$$J(k, t_k) = \frac{m_{\text{gauche}}}{m}\text{MSE}_{\text{gauche}} + \frac{m_{\text{droit}}}{m}\text{MSE}_{\text{droit}}$$

Avec $m_{\text{gauche/droit}}$ étant le nombre de données dans le nœud à gauche ou à droite, m et le nombre de données total fournies à l'arbre, $\text{MSE}_{\text{gauche/droit}}$ est l'erreur quadratique entre la valeur estimée $\hat{y}$ et la valeur réelle y_i. L'algorithme itère à travers toutes les combinaisons possibles, ce qui donne une complexité de $\mathcal{O}(e^m)$, le problème de trouver l'arbre optimal est un problème NP-complet, l'algorithme trouve une assez bonne solution au problème qui n'est pas à priori la solution (l'arbre) optimal.

Les arbres de décisions sont une bonne solution pour les problèmes de régression non linéaire, cependant, ces arbres sont très instables, car ils s'adaptent aux données sans contrainte et peuvent facilement causer des problèmes de sur-ajustement (overfitting), de plus, les arbres de décisions ont une variance élevée, leur entraînement sur deux échantillons du même jeu de donnée peut donner des résultats très différents. En pratique, il faudra choisir délicatement les hyper-paramètres du modèle (profondeur, taille d'échantillon par nœud, pénalité, etc) pour éviter ces situations et imposer des contraintes au modèle.

Random Forests : c'est un algorithme de Machine Learning révolutionnaire et basé sur les arbres de décisions, développé initialement par Ho (1995). L'idée se base sur le concept du **bootstraping** en statistique, une technique qui permet d'estimer une statistique d'une manière fiable en le calculant par rapport à n échantillons tirés avec remplacement du jeu de données. Entraîner plusieurs modèles du même type sur des échantillons de données à travers le bootstrapping et agréger par la suite leur résultat dans un seul modèle permet d'obtenir des résultats plus fiables, plus précis avec moins de biais et de variance ($\frac{\sigma}{n} < \sigma$). Cette approche s'appelle l'apprentissage par ensemble (ensemble learning) et l'algorithme de Random Forest est basé sur ce principe. (Voir annexe B pour découvrir les différentes étapes de cet algorithme).

Gradient boosted trees : un autre algorithme puissant de régression et basé sur l'approche d'apprentissage par ensemble est l'algorithme du gradient boosted decision trees (GBDT). Cet algorithme entraîne un ensemble B d'arbres de décisions sur les résidus (gradient d'erreur) de chaque arbre de manière séquentielle, ce qui assure une réduction continue de l'erreur. (Voir annexe B pour découvrir les différentes étapes de cet algorithme).

Apprentissage non-supervisé (unsupervised learning)

L'apprentissage non supervisé (unsupervised learning) est un autre type d'apprentissage machine, dans ce cas, on dispose que des données d'entrée X et pas de la variable observé y. Ce type d'apprentissage se base principalement sur du clustering et détecter des motifs communs dans les données d'entrée X. Il est utilisé pour effectuer du clustering comme l'algorithme de **k-means** ou pour détecter des anomalies dans les données, par exemple, développer un modèle capable de détecter des transactions frauduleuses.

L'apprentissage non supervisé est aussi utilisé récemment pour générer des données telles que des images, des vidéos, de la voix, etc.

Apprentissage profond (deep learning)

Un type d'apprentissage particulier et très populaire est l'apprentissage profond (deep learning). Cette approche à l'apprentissage est purement inspirée du cerveau humain et se base sur les réseaux de neurones artificiels.

Un réseau de neurones artificiels est composé de plusieurs couches, chaque couche est elle composé d'une unité de calcul basique appelée un neurone artificiel, ce neurone reçoit en entrée une combinaison linéaire des donnée calculée par la couche précédente et applique une entre transformation ainsi qu'une fonction d'activation g_i qui est similaire au principe d'activation des neurones humains par un signal électrique dépassant un certain seuil.

Le but d'un réseau de neurones artificiels est d'approximer une fonction f quelconque (non linéaire ou linéaire) pour des problèmes de classification, de régression et dans certains cas d'autre type de tâches comme la génération de contenu. On parle donc d'approximateur universel, car en théorie il est possible d'estimer n'importe quelle fonction f^* à travers un réseau de neurones. Le nombre de couches et le nombre de neurones dans chaque couche ainsi que d'autres paramètres déterminent la précision d'approximation. Ceci permet de modéliser des relations non linéaires en transformant les données d'entrée x vers $\phi(x)$. Si y est donnée par la relation :

$$y = f(x, \theta, w) \tag{2.4}$$

Alors, il est possible de l'écrire de cette manière :

$$y = \phi(x, \theta)^t w \tag{2.5}$$

Le but du réseau de neurones est donc d'estimer cette transformation ϕ en estimant les différents poids $w = (w_1, w_2, w_3, .., w_n)^t$ dans chaque neurone.

Note : θ est un vecteur d'hyper-paramètres déterminé par l'utilisateur du réseau avant l'apprentissage.

L'apprentissage dans ces réseaux s'effectue à base de gradient et d'une fonction de coût : La fonction de coût $J(w)$ permet de calculer une erreur selon les paramètres et les poids du réseau. Cette fonction de coût doit être lisse et dérivable, un exemple d'une fonction de coût est la somme des carrés d'écarts qui est lisse et dérivable :

$$J(w) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{f}_w(x^{(i)}; \theta) - y^{(i)})^2 \tag{2.6}$$

L'apprentissage s'effectue par la suite en calculant le gradient de la fonction de coût $\nabla_w(J)$ par rapport aux poids, de prendre son négatif et calculer ensuite les nouveaux poids w_i à l'itération k via un pas γ selon la règle suivante (écrite sous la forme vectorielle) :

$$w_k = w_{k-1} - \gamma \nabla J(w_n) \tag{2.7}$$

Cet algorithme permet après la fin d'un certain nombre d'itérations de trouver les poids w optimaux qui convergent vers un minimum local de la fonction de coût et donc donne une bonne estimation de la fonction $\phi(x; \theta)$.

Il existe plusieurs types de réseaux de neurones artificiels comme les réseaux simples (feed forward networks) employé pour la régression ou la classification, les réseaux récurrents (RNNs) utilisé dans le traitement de langage naturel et sur des données de nature séquentiel, les réseaux à convolution (CNNs) utilisé principalement sur les données sous forme d'images ou de vidéos ainsi que les réseaux génératifs (VAEs, GANs, etc) qui permettent de générer de nouvelles données comme des images, des vidéos ou autre (voir l'annexe C pour plus de détails sur les réseaux de type VAE).

2.2.4 Synthétisation de données

La synthétisation de données est le processus de générer des données synthétiques (artificielles) qui ont les mêmes propriétés que les données réelles. Les données synthétiques sont utilisées pour améliorer les résultats des différents algorithmes d'apprentissage. Cette approche est employée de plus en plus dans des contextes où la récolte de données réelles est compliquée ou parfois impossible, cette approche est de plus en plus populaire, il est estimé qu'en 2024 environ 60% des données employées pour entraîner des modèles intelligents se basera sur des données synthétiques (Castellanos, 2021) (Gartner, 2021). La synthétisation de données permet de générer une quantité importante de données qui peuvent améliorer les performances des modèles de Machine Learning sans pour autant les biaiser, car ces données sont basées sur les propriétés des données réelles.

Bien qu'il soit possible de générer des données sous formes d'images, de vidéos, audio et autres formats, en ce qui suit, nous allons nous intéresser aux données tabulaires, ce sont des données structurées et se divisent en deux types : Numérique et catégorique.

- Les données numériques : ce sont les données qui peuvent être continues comme le poids, la longueur, etc. Ou des données discrètes comme le nombre de conteneurs par jour.
- Les données catégoriques : qui eux se divisent en deux : ordinales comme les jours de la semaine ou nominales comme la couleur.

La figure 2.11 résume les types des données tabulaires :

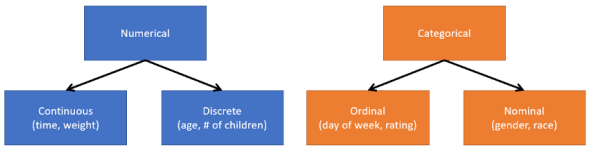


FIGURE 2.11 – Les différents types de données tabulaires (Unzueta, s. d.)

Le problème de synthétisation de données consiste à utiliser un jeu de données tabulaire T qui contient N_d colonnes discrètes et N_c colonnes continues et de générer des données tabulaires est d'entraîner un réseau générateur G qui apprend à générer un jeu de données synthétique T_{synth} depuis le jeu T.

Il existe principalement 3 approches pour synthétiser des données et plusieurs méthodes dans chaque approche :

À travers des méthodes statistiques :

- **La méthode de Monte-Carlo :** l'approche de Monte-Carlo est une méthode statistique classique inventé vers la fin des années 1940 par l'informaticien John Von Neumann et le mathématicien Stanislaw Ulam , Cette méthode consiste à échantillonner des données depuis un ensemble de combinaison possible connus suivant une distribution prédéfinie tel qu'une loi normale avec des paramètres connus, c'est donc une méthode paramétrique. L'avantage de cette méthode est qu'elle est facile à implémenter et efficace en termes de calcul, cependant ces résultats ne sont pas très précis, car elle suppose que les données suivent une loi de distribution avec des paramètres connus au préalable, ce qui est rarement le cas en réalité.
- **Le boostrap :** le bootstrap est une méthode statistique classique, elle permet de calculer la distribution d'une statistique comme une moyenne, un écart-type ou autres mesures en échantillonnant les données avec remplacement. À chaque itération, nous obtenons un nouveau jeu de données, on peut fusionner ces jeux de données et obtenir des données synthétiques. L'avantage du bootstrap comparé à la méthode de Monte-Carlo est que le bootstrap est une méthode non paramétrique, il ne suppose aucune distribution au préalable et utilise la distribution des données réelles pour générer de nouvelles données. Néanmoins, cette méthode reste incapable à prendre en considération la corrélation entre deux variables (prédicteurs/colonnes) ou plus.

À travers la simulation à événements discrets :

Cette approche est très fréquente quand le processus qui génère les données est maîtrisé, ce qui est le cas dans plusieurs applications industrielles. Le principe est de modéliser tout le processus étudié à travers un logiciel de simulation comme AnyLogic (voir annexe

À travers des réseaux de neurones artificiels (deep learning) :

Le problème de synthétisation de données à travers le deep learning est un domaine de recherche d'actualité, les réseau de neurones, combinés à des technique statistique ou de l'apprentissage non supervisé permettent de générer des données réalistes et très proches des données originales, plusieurs modèles ont été mis en place ces dernières année (entre 2014 et 2022), nous allons présenter les 3 grandes familles :

- **Tabular Variational Auto-Encoder (VAE) :** un algorithme non supervisé qui peut apprendre la distribution d'un ensemble de données original et générer des données synthétiques via une double transformation, une première transformation à travers un réseau qui encode les données dans un espace avec des dimensions plus petites que les dimensions initiales (un espace latent) et un décodeur qui essaye de reconstruire la donnée originale depuis l'espace latent. Le modèle ajoute une erreur de reconstruction avant le décodage, qui peut être minimisée par un apprentissage itératif. Ce processus permet d'entraîner un réseau à deux blocs (un encodeur et un décodeur) qui peut générer des données (Voir annexe C pour plus de détails sur leur fonctionnement).
- **Conditional Tabular Generative Adversarial Networks (CTGAN) :** Un réseau de neurones génératif développé par Xu et al. (2019b) et inspiré de l'architecture des GANs (Generative Adversarial Networks) (Goodfellow et al., 2014). Le principe est d'entraîner un générateur G et un classificateur (appelé discriminateur D), le générateur essaie de générer des données aussi proches de la réalité afin de

tromper le classificateur D. Ce processus permet de générer des données très réalistes et très semblables à la réalité. Les CTGANs sont une variation de ce réseau adapté aux données tabulaires.

— **Modèles de diffusion :** Un type de réseau de neurones inspiré de la thermodynamique et qui se base sur l'apprentissage non supervisé (unsupervised learning) (Sohl-Dickstein et al., 2015). Ce type de modèle prend en entrée une donnée et ajoute une erreur gaussienne à cette dernière, à chaque itération le réseau rajoute une erreur plus importante jusqu'à ce que la donnée soit détruite et devient identique à un bruit aléatoire. Par la suite, le réseau essaiera de reconstruire la donnée initiale. Cet apprentissage lui permet de générer des données réalistes dans plusieurs formats.

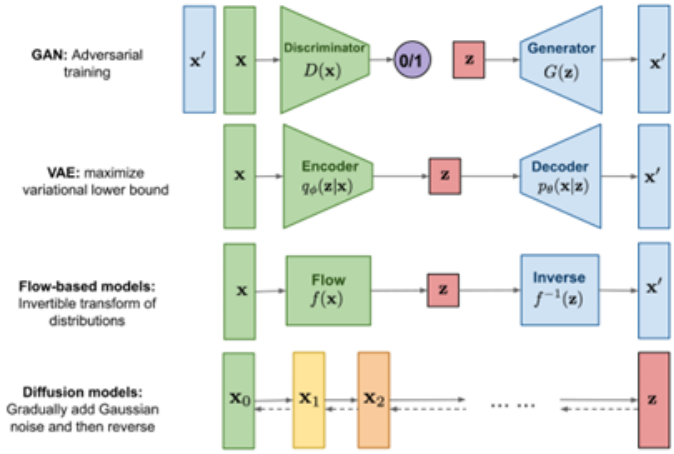


FIGURE 2.12 – Architecture des différents modèles de synthétisation des données (Weng, 2021)

Maintenant qu'on a défini l'ensemble des outils nécessaire pour la résolution de la problématique, nous allons passer à l'utilisation de ces outils dans l'estimation des lead times d'importation.

2.3 Détermination des lead times dans les opérations d'importation à travers la BI et la Data Science

Après avoir présenté les outils nécessaires pour résoudre notre problématique, nous allons désormais détailler l'utilisation de ces outils dans l'estimation des lead times et par la suite donner une formulation de notre problématique.

L'estimation des lead times dans les opérations d'importations est possible via plusieurs outils dont la Business Intelligence qui facilite la compréhension des différents processus et la détermination des performances de différentes parties prenantes. Les techniques de Machine Learning quant à elles sont utilisées pour mettre en place un modèle de prévision des différents lead times. Et dans le cas où les données existantes ne suffisent pas pour entraîner des modèles de prévision, les techniques de synthétisation des données peuvent être utilisées pour générer plus de données et améliorer la précision du modèle de prévision.

2.3.1 Business Intelligence pour l'analyse des performances

Un objectif important pour plusieurs entreprises est l'accès à l'information correctement et rapidement, les solutions de Business Intelligence permettent de réaliser cet objectif. La Business Intelligence donne un état des lieux agrégé et en détail à la fois des performances de l'entreprise, de ses fonctions et de son environnement.

Plus précisément, la Business Intelligence est utilisée en management de la Supply Chain pour extraire des informations et du savoir à partir des données récoltées, elle présente la partie descriptive du processus de Supply Chain Analytics car elle permet de décrire le passé et le présent à travers les données. La planification de la supply chain est une étape qui peut profiter d'une bonne intégration de la Business Intelligence : en analysant les performances, nous pouvons planifier la bonne configuration de la supply chain qui minimise les coûts et lead times. Avec le développement des nouvelles technologies comme Internet of Things (IoT), le cloud et la 5G, la Business Intelligence en temps réel devient de plus en plus accessible, ceci permettra aux entreprises de développer des supply chains plus réactives et intégrer davantage le processus de Supply Chain Analytics dans les fonctions de management afin de prendre des décisions bien fondées. Azvine et al. (2005) ainsi que Nguyen et al. (2005) étudient les inconvénients des systèmes traditionnels de la BI et proposent des architectures potentielles pour la Business Intelligence en temps réel.

En particulier, notre travail emploie la Business Intelligence comme une approche d'analyse des performances concernant l'importation, et cela, à travers de tableaux de bord en temps réel, la dimension de temps réelle est intégrée dans notre cas à travers la plateforme Microsoft Sharepoint qui servira comme source de données dynamique en temps réelle, le tableau de bord donc peut changer automatiquement à chaque fois que les données sont modifiées sur MS Sharepoint.

2.3.2 Estimation des lead times avec le Machine Learning

L'estimation des différents délais (lead times - LT) dans les processus industriels est une tâche importante dans la planification de la supply chain et la prise de décision, ce problème a souvent été traité à travers des modèles déterministes qui ne prennent pas la nature stochastique des lead times. Ioannou et Dimitriou (2012) ont proposé un algorithme déterministe et itératif qui permet d'estimer les lead times à travers le MRP (Material Requirements Planning), à travers la méthode du Kanban (Weflen et al., 2022), à travers le savoir-faire de l'entreprise (Mourtzis et al., 2014) ou par d'autres méthodes analytiques (Gyulai et al., 2018).

Une approche plus moderne, qui fera l'objet de ce travail, est l'utilisation du machine learning comme une méthode plus robuste pour l'estimation des lead times des processus industriels, ceci permet de prendre en considération la nature stochastique du problème comme montré précédemment par les algorithmes de machine learning (Cheng et al., 2020) ;(Gyulai et al., 2018) ;(Lingitz et al., 2018). D'autres travaux ont employé le machine learning avec le process mining (Welsing et al., 2020) ou encore l'approche CRISP-DM (Cheng et al., 2020) qui sera utilisé dans ce travail comme un cadre pratique pour développer notre solution.

Ces travaux s'intéressent principalement à l'estimation des lead times de fabrication, tandis que nous allons nous intéresser aux lead times de l'importation au sein d'une entreprise de services pétroliers. Le lead time totale d'importation est composée de plusieurs lead times (présentés dans le chapitre 3), il existe plusieurs travaux dans la littérature

qui traitent l'estimation des lead times de livraison et d'arrivage en général : Hayya et al. (2013) étudient les lead times comme une variable aléatoire dans la livraison juste à temps, Tsolaki et al. (2022) trouvent qu'une grande partie des applications du Machine Learning en supply chain s'intéressent à l'estimation du temps d'arrivée. Notre travail peut être abordé sous cet angle, car le lead time totale d'importation est considéré comme un délai d'arrivée plus complexe puisqu'il concerne plusieurs parties prenantes.

2.3.3 Formulation de la problématique

Comme présenté dans le premier chapitre, les lead-times d'importation dans les supply chains globales des services pétroliers représentent un facteur important de leurs performances. L'estimation de ces lead-times permet de mieux planifier ces supply chains et d'anticiper les retards ce qui influence sur la performance financière et opérationnelle de ces supply chains.

Le but de notre travail est d'estimer les lead times reliés au processus d'importation d'une marchandise dans le cadre d'une entreprise de service pétrolier. Notre étude s'intéresse seulement aux lead times qui comportent les étapes de dédouanement en Algérie ainsi que la livraison d'expédition vers les bases opérationnelles. Les lead times au niveau international (livraison aux hubs (centres de distribution) et autres) sont déjà calculés et ne seront pas traités par ce travail. Par la suite, nous utiliserons ces estimations pour calculer le cout de stockage et de livraison au niveau national pour chaque expédition.

Pour cela, nous utiliserons la Business Intelligence pour analyser les performances des opérations d'importation au niveau de l'entreprise concernée. Nous allons ensuite développer des modèles d'estimation du lead time et les comparer. Cependant, les modèles de Machine Learning nécessitent un volume important de données, et afin de remédier à ce problème, nous allons synthétiser des données à travers deux méthodes : la simulation et les réseaux de neurones artificiels de type TVAE (Tabular Variational Auto-Encoders).

Chapitre 3 : État des lieux

Chapitre 3

État des lieux

Après avoir exploré les différents concepts théoriques clés liés à notre problématique dans les deux premiers chapitres, nous allons à présent mettre en place le cadre d'application de ce travail.

D'abord on présentera le marché des services pétroliers et son évolution, par la suite nous allons présenter l'entreprise Schlumberger et ses filiales (Business Lines), pour finir nous allons explorer la supply chain global de l'entreprise et nous nous focaliserons sur son processus d'importation et d'exportation.

3.1 Marché des services pétroliers

Le marché des services pétroliers est composé des entreprises qui offrent des services et des produits relatifs à l'extraction du pétrole et des gaz ainsi que leur exploitation. En 2020, la taille de ce marché a été estimée à 96.65 milliards de dollars américains qui pourrait atteindre une valeur de 135 milliards de dollars américains en 2027 (Intelligence, 2022), cette croissance est principalement menée par la demande croissante des hydrocarbures ainsi que la découverte de nouveaux gisements en offshore. Les services proposés sont divers, mais les services de production et de forage dominent ce marché, de plus le taux de croissance du marché dépend fortement de la région : en Nord d'Amérique on retrouve une croissance élevée de 18.8% en 2001 à 25% en 2020, car dans cette région, les gisements de pétrole sont plus durs à exploiter à cause de leur nature, ce qui implique une demande croissante des services pétroliers, notamment les services de forages (Intelligence, 2022).

En Asie, en Europe, en Amérique latine et en Océanie on retrouve une croissance modérée, tandis qu'en Afrique et qu'au Moyen-Orient, la croissance est plus faible à cause des ressources déjà exploitées depuis plusieurs années et la volatilité des prix du pétrole, la demande est donc constante ou évolue doucement.

FIGURE 3.1 – Croissance du marché des services pétroliers entre 2021-2022 par région (Intelligence, 2022)

Parmi les principaux acteurs du marché des services pétroliers on retrouve : Schlumberger, Baker Hughes, Halliburton, General Electric, DMC Global et autres. Certaines entreprises technologiques prennent part aussi dans ce secteur tel que Siemens ou encore le groupe ABB.

L'évolution de ce marché est corrélée avec l'évolution du marché pétrolier, qui entre 2019 à 2022 a connu une baisse importante dans la production et une forte volatilité des prix. L'Algérie, étant un pays ou les hydrocarbures représentent plus de 80% des exportations, a été impacté économiquement par cette pandémie, durant les 9 premiers mois de 2020, les prix du pétrole Sahara Blend ont chuté de 39.5%, la contribution des revenues des hydrocarbures au PIB du pays pourrait aussi diminuer de 13% à 9% à cause de la pandémie (World Bank Group, 2020). Le marché des services pétroliers en Algérie à son tour a été impacté économiquement et organisationnellement, plusieurs entreprises du secteur telles que Schlumberger Algérie ou GE Algérie ont été obligées d'arrêter leurs activités pendant une certaine période infligeant par cela des coûts importants.

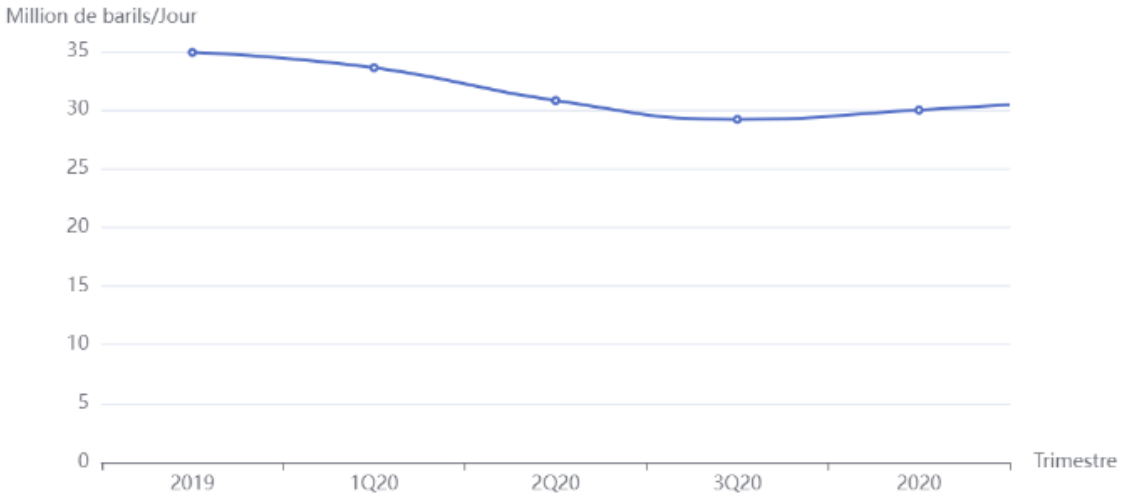


FIGURE 3.2 – Évolution des prix du pétrole pour les pays membre de l'OPEC durant la pandémie (*Oil 2021*, 2021)

Après avoir présenté le marché des services pétrolier global et par région ainsi que son évolution, nous allons à présent se focaliser sur l'entreprise Schlumberger, qui est classée comme première entreprise de services pétroliers à travers le monde.

3.2 Présentation de Schlumberger

Notre travail s'inscrit dans la supply chain de Schlumberger, il est donc important de présenter cette entreprise ainsi que son organisation et son histoire au niveau globale, nous allons par la suite présenter Schlumberger Algérie en détail.

Schlumberger est une multinationale franco-américaine fondée en 1926, dont la principale mission est de fournir les technologies, les services de gestion intégrée de projets, et des solutions d'information pour l'industrie de l'exploration et l'exploitation pétrolière et gazière internationale, elle opère sur un ensemble de 120 pays.

La fiche suivante résume les informations essentielles de l'entreprise :

Schlumberger	
Date de création	1926
Fondateurs	Conrad Schlumberger et Marcel Schlumberger
Forme juridique	Société anonyme avec appel public à l'épargne
Siège sociale	Bureaux principaux à Houston (USA), Paris (France), et la Haye (Pays-Bas).
Direction	CEO : Olivier Le Peuch EVP et CFO : Stephane Biguet
Secteur d'activité	Service pétrolier
Effectif	92000 de 160 nationalités (2021)
Capital sociale	57 000 millions USD (2021)
Chiffre d'affaire	22 930 millions USD (2021)
Résultat net	188 millions USD (2021)

TABLE 3.1 – Informations utiles sur Schlumberger (Schlumberger Ltd, 2021).

3.2.1 Organisation de Schlumberger

Aujourd'hui, Schlumberger opère sur 120 pays et afin de maintenir ses performances, elle est divisée en cinq bassins géographiques (les Amériques, l'Asie, la Russie et l'Asie centrale, le Nord d'Afrique et le Moyen-Orient, L'Offshore de l'Atlantique) qui partagent des besoins technologiques similaires, chaque bassin géographique est divisé en GeoUnit, il existe au total 30 GeoUnits. Une GeoUnits est un pays ou un groupe de pays gérés dans l'un des cinq bassins. La carte ci-dessous représente la répartition des bassins ainsi que les GeoUnits de Schlumberger :

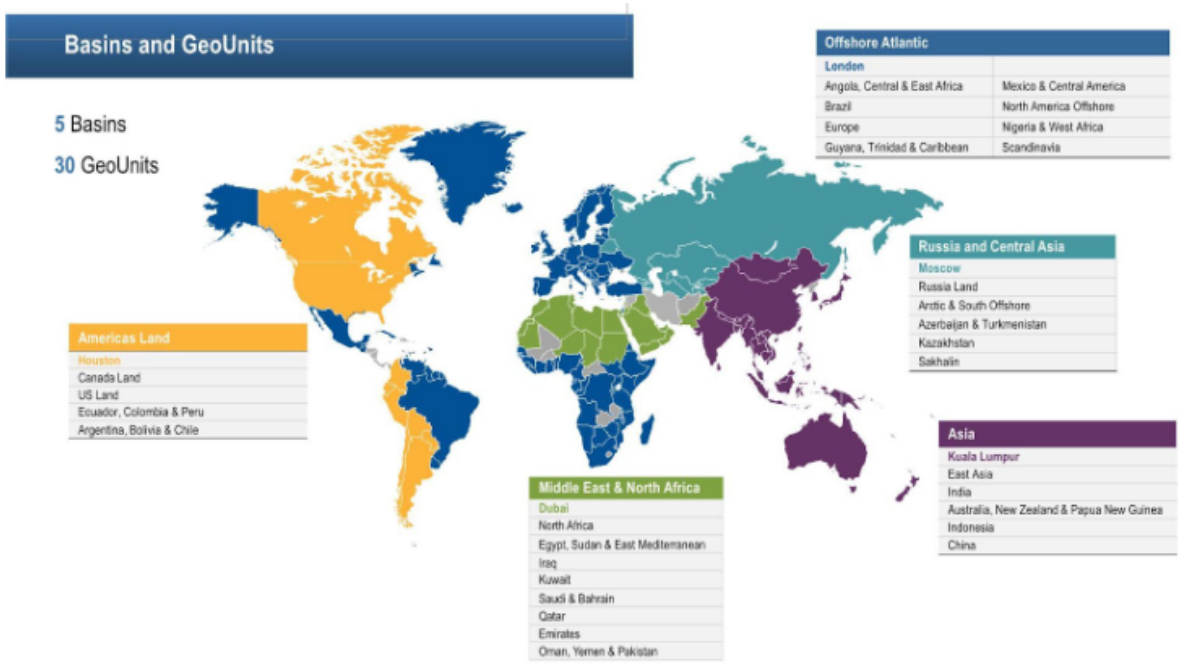


FIGURE 3.3 – Carte des bassins et des GeoUnits de SLB Source : documents internes à Schlumberger

Afin de mieux gérer ses activités, SLB est structurée en divisions et chaque division englobe plusieurs Business Lines, SLB compte exactement 4 divisions qui sont : Digital and Integration, Reservoir Performance, Production Systems et Well Construction. Nous allons présenter brièvement chaque division :

- **Digital Integration :** la division Digital Integration (DI) intervient dans la récolte, l'étude et l'analyse des données sismiques et géologiques.
- **Reservoir Performance :** la division Reservoir Performance (RP) se base sur les technologies et services qui garantissent l'optimisation et la performance des réservoirs.
- **Production Systems :** la division Production Systems (PS) stimule l'innovation technologique et l'intégration totale du système, de l'interface réservoir-puits à mi-chemin.
- **Well Construction :** cette division intègre des activités, processus et de la technologie de forage intelligent qui maximisent la précision, l'efficacité et la valeur, et minimisent les risques, tout en assurant des changements dynamiques.

3.2.2 Schlumberger en Algérie et son organisation

Schlumberger Ltd. se présente en nord d'Afrique à travers sa division géographique Schlumberger NAF (North Africa GeoMarket) qui regroupe cinq pays : Algérie, Maroc, Tunisie, Libye et le Tchad. Ce GeoMarket représente un chiffre d'affaires important compte tenu de la richesse de la région en pétrole et gaz naturel répartie sur l'ensemble du Sahara à environ 898 7000 barils de pétrole produite par jour en 2020 en Algérie (un chiffre à la baisse due à la pandémie) (OPEC, 2021).

Le manager de Schlumberger NAF travaille en coordination avec les managers de chaque pays de cette division géographique, ainsi que les managers des fonctions support : ressources humaines (HR), risque et sécurité (HSE), supply chain (SC), etc.

FIGURE 3.4 – Schlumberger North Africa GeoUnit, source : documents internes à Schlumberger

En Algérie, l'entreprise est installée en 1955 et possède un siège au niveau de la zone d'activités de Chéraga, route d'Ouled-Fayet à Alger qui représente aussi le siège social du "North Africa GeoMarket", elle possède aussi un ensemble de 11 bases opérationnelles au niveau des 4 zones d'activités : Hassi Messaoud, Ain Amenas, Hassi Berkine et Ain Salah.

En plus de ces bases opérationnelles, l'entreprise possède aussi des bases logistiques, des bunkers d'explosifs et des Guest House (maison d'hôtes).

Elle est présente sur le pays à travers deux entités légales : COPS (Compagnie des Opérations Pétrolières Schlumberger) et SPS (Services Pétroliers Schlumberger) et elle fournit des services pétroliers principalement à l'entreprise nationale SONATRACH ainsi qu'aux entreprises : Total, Anadarko, British Petroleum, AGIP et autres. Parmis ses services en Algérie :

- L'installation des bases opérationnelles.
- Les études géologiques et sismiques.
- La construction des puits.
- Le test des puits.
- L'importation des équipements requis.

Maintenant que nous avons présenté la structure globale de Schlumberger, nous allons détailler le fonctionnement de sa supply chain au niveau international et national, ce qui nous permettra par la suite d'aborder le processus d'importation et d'exportation.

3.3 La Supply Chain de Schlumberger

La supply chain au niveau de Schlumberger représente une fonction de support importante qui impacte directement les performances de toute l'entreprise, cette fonction fait partie

de “Shared Services Organization - S.S.O” qui regroupe l’ensemble des 7 fonctions de support : “HR Support”, “Finance”, IT Operations, “Procurement Sourcing”, “Contracts”, “Distribution” et “Construction Facilities” au sein de Schlumberger qui coordonnent entre les différentes Business Lines.

La supply chain au sein de Schlumberger vise à permettre aux différentes Business Lines de satisfaire leur besoin en termes de ressources en matériels à travers un réseau logistique qui permet de transférer des équipements et du matériels entre ses bases, hubs et autres sources à travers le monde avec un coût optimal.

3.3.1 Organisation de la supply chain international chez Schlumberger

Schlumberger étant une entreprise multinationale présente sur 120 pays a développé au fil du temps une supply chain globale qui lui permet de s’approvisionner de plusieurs fournisseurs à travers le monde et de garder des stocks dans des endroits stratégiques lui permettant de rester réactif. Dans cette partie, nous allons présenter les éléments qui constituent cette supply chain globale.

Le design de la supply chain de Schlumberger au niveau international a évolué au fil des années. L’entreprise a établi un modèle de hubs GOLD (Global Oilfield Logistics and Distribution) en 1999 qui consiste à établir des centres de services distributions (DSC).

Ces centres de services de distributions (Distribution Service Center - DSC) sont des hubs départagés stratégiquement à travers le monde afin de fournir des localisations communes à un ensemble de Géo Units/GeoMarkets pour consolider les commandes, stocker les produits et les équipements demandés fréquemment ou les produits critiques. La figure suivante illustre la répartition des DSC de Schlumberger à travers le monde.

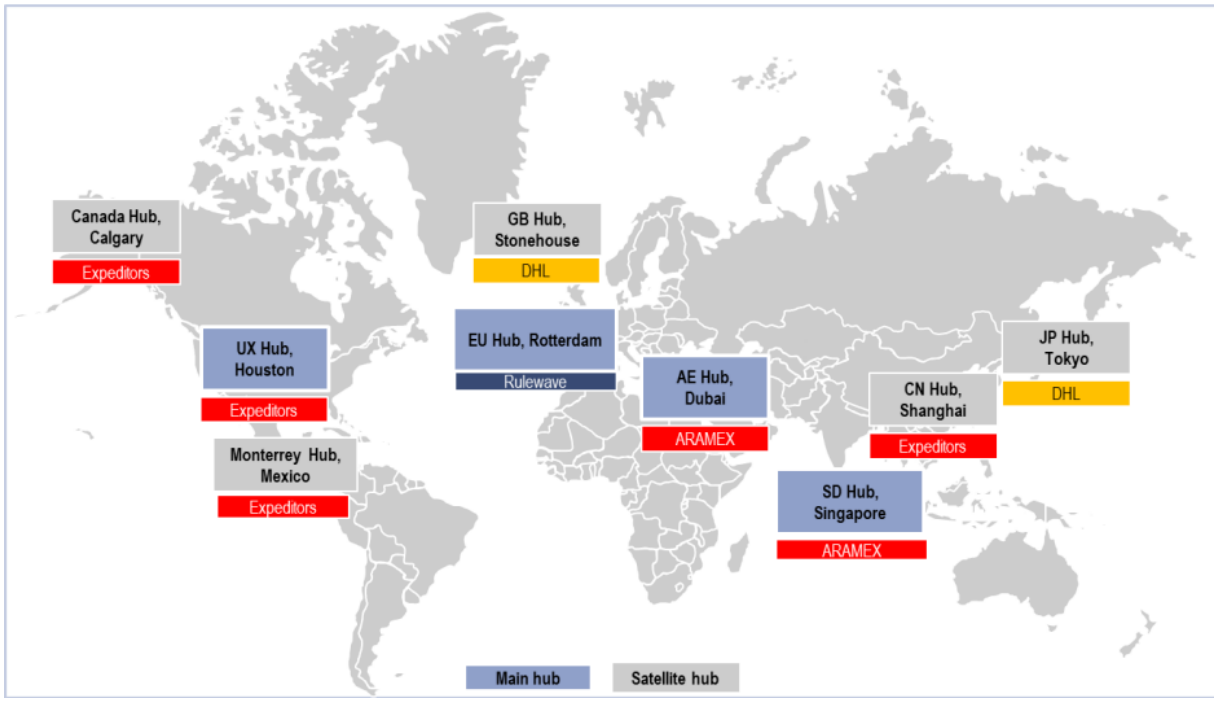
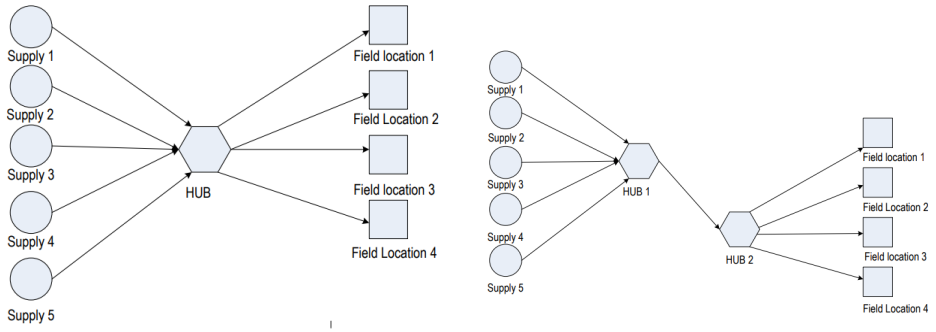


FIGURE 3.5 – la répartition des DSC de Schlumberger à travers le monde, source : documents internes

Ce réseau logistique est un modèle “*Hub and Spoke*”. Il peut se présenter sur un niveau

pour une seule consolidation ou deux niveaux pour deux consolidations (Figure 3.6) :



(a) Avec une seule consolidation (b) Avec deux consolidations

FIGURE 3.6 – Modèles "Hub and Spoke" de Schlumberger avec un niveau et deux niveaux de consolidation (Document interne à Schlumberger).

Le spécialiste Import/Export peut opter pour une expédition GOLD qui passe par les DSC (les hubs logistiques) et est consolidée à un niveau ou deux niveaux, ce type d'expédition est le cas par défaut et représente plus de 70% des expéditions de Schlumberger Algérie.

Une expédition non-GOLD est une exception à la règle et concerne principalement les commandes urgentes ou les commandes effectuées chez des fournisseurs non GOLD (des commandes rares) et donc ne passent pas par les centres de distribution (DSC) et sont livrées directement à la GeoUnit.

3.3.2 La supply chain chez Schlumberger Algérie

La supply chain de Schlumberger en Algérie est constituée des trois départements suivants :

- **Achats & Approvisionnement (PS) :** ce département est composé de deux fonctions : l'approvisionnement qui a pour but de trouver les fournisseurs et établir des relations avec eux et l'achat qui travaille sur l'établissement des contrats avec ces fournisseurs.
- **Global Distribution :** elle est constituée de 3 fonctions principales : l'import-export, le material management qui s'occupe de la bonne circulation et gestion du flux matériels et l'inventaire de l'entreprise.
- **Domestic Transport :** cette fonction s'assure de la bonne circulations des équipements et des différentes ressources entre les bases opérationnelles et les chantiers de travail.

Après avoir présenté la supply chain globale de Schlumberger ainsi que son implémentation en Algérie, et afin de mieux cerner la problématique étudiée, nous allons maintenant présenter le processus d'importation et d'exportation de Schlumberger.

3.4 L'importation et l'exportation chez Schlumberger

Afin de mieux gérer la fonction d'importation et d'exportation, Schlumberger a standardisé le processus en mettant en place des procédures claires pour chaque régime d'importation, en précisant les tâches à réaliser et ses échéanciers ainsi que les responsables de chaque tâche. Un aperçu du standard se trouve en annexe

3.4.1 Fonctionnement du processus d'importation et d'exportation

Le processus englobe quatre grandes phases :

1. **Expression et validation du besoin :** l'équipe de Schlumberger définit le besoin exact à travers l'ERP SAP, après que ce besoin est validé par les supérieurs, on passe à la phase de planification et d'achat.
2. **Planification et achats :** l'équipe du Logistic Control Tower (LCT) va satisfaire cette demande via une des deux méthodes :
   — Un transfert, dans le cas où l'équipement est disponible en stock ;
   — Dans le cas échéant, un bon de commande (Purchase Order) doit être lancé, et le fournisseur choisi sera notifié pour préparer et expédier la commande aux entrepôts des HUB de SLB.
3. **Distribution :** La commande va maintenant passer par deux étapes de distribution : à l'international et au niveau national.
4. **Le dédouanement :** ce processus important englobe plusieurs parties prenantes qui sont : les transitaires, les spécialistes I/E et les services de douane détaillés dans le premier chapitre.

   Ce processus est initié avec un Green Light donné par le chargé d'I/E. Après le démarrage d'une expédition, l'équipe Schlumberger aura le ETA (Estimated Time of Arrival), c'est une date importante qui permet aux chargés d'I/E et les transitaires de mettre en place un planning et préparer l'ensemble des documents nécessaires. Ce processus englobe six tâches :
   — **I/E Contrôle :** le chargé d'I/E doit vérifier le HTC (Harmonized Tariff Code), et émet une demande d'autorisation dans le cas où l'expédition nécessite une.
   — **Déclaration douanière :** Le transitaire prends le relai, et il soumet le dossier au bureau de douane, le dossier est constitué de :
      — La facture d'achat signée et cachetée.
      — Lettre de transport aérien pour le fret aérien ou lettre de transport maritime.
      — Autorisation dans le cas où l'expédition nécessite une autorisation.
      — Attestation et engagement (attestation d'existence de l'entreprise).
   — **Paiement des droits et taxes :** Le transitaire devra vérifier ensuite payer les droits et taxes, il sera payé par Schlumberger plus tard
   — **libération douanière :** Dans certains cas, la douane impose une visite de l'expédition pour vérifier qu'il s'agit vraiment du produit déclaré, ensuite l'expédition est libérée avec la remise d'un bon à enlever.
   — **Livraison :** cette étape est assurée par le transitaire, elle englobe plusieurs étapes :
      — Soumission des bons de livraison.
      — Facturation des frais de stockage.
      — Paiement des frais de stockage.
      — Réservation et chargement du camion.
      — Déchargement dans la base SLB.
   — **Paiement :** une fois que l'expédition est dans la base de Schlumberger, le transitaire émet une facture avec tous les coûts.

3.4.2 Les différents lead times et coûts liés à l'importation

Un lead time est la durée nécessaire pour effectuer une tâche, le processus d'importation regroupe plusieurs étapes et chaque étape prend un certain lead time. On peut classer les lead times selon la partie prenante concernée : Une business line au sein de Schlumberger ou le CCA (Customs Clearance Agent), les services douaniers.

Le tableau **??** donne une définition des dates utilisée par Schlumberger, tandis que le tableau 3.3 présente chaque lead time ainsi que l'entité concernée par ce dernier :

Date	Définition
CFD : Complete File Date	Date d'envoi du dossier complet de déclaration de la part business line au CCA
Declaration Date	Date de déclaration de l'expédition de la part du CCA au niveau de la douane
VAT Submission Date	Date d'envoi de la demande de remise sur la TVA (VAT Exemption) (valable que pour les expéditions sous régime VAT).
VAT Recovery Date	Date de réception de la remise sur la TVA pour une expédition (valable que pour les expéditions sous régime VAT).
Release Date	Date de remise du bon à enlever, ce qui permet à l'expédition de sortir de l'entrepôt douanier
Delivery Date	Date de début de la livraison de l'expédition par le transporteur du CCA de l'entrepôt douanier vers la base opérationnelle.
Reception Date	Date de réception de l'expédition au niveau de la base opérationnelle
Restitution Date	Date de remise du conteneur vers le port d'entrée (valable seulement pour les expédition de type Sea Freight

TABLE 3.2 – Les différentes dates relatives à l'importation chez Schlumberger

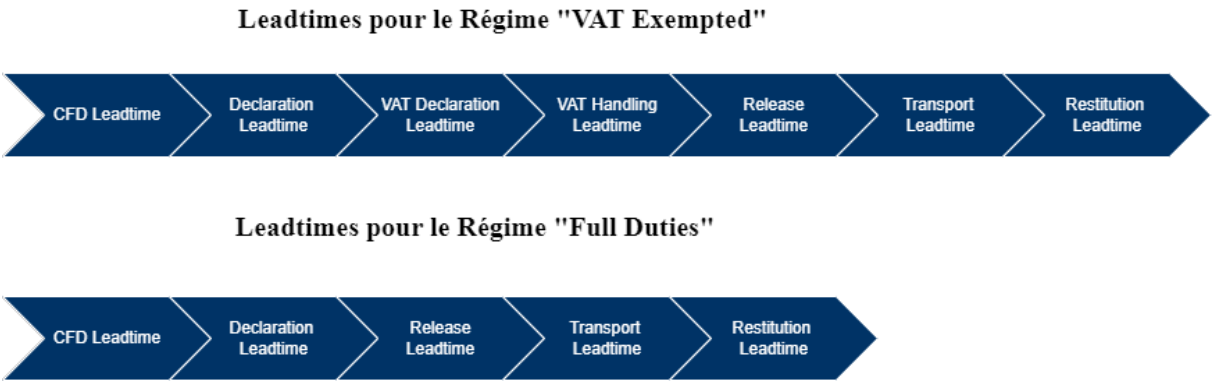


FIGURE 3.7 – Les différents lead times pour les deux régimes d'importation permanente.

Lead Time	Définition	Calcul du lead time	Entié
CFD Lead Time	Durée nécessaire pour qu'une business line prépare les documents de déclaration et les envois au CCA	CFD - ATA	Business Line
Declaration Lead Time	Durée nécessaire pour que le CCA déclare l'expédition aux autorités douanières.	Declaration Date - CFD	CCA
VAT Declaration Lead Time	Durée que prend le CCA pour soumettre une demande de remise sur la TVA (pour le régime VAT)	VAT Submission Date - Declaration Date	CCA
VAT Handling Lead Time	Durée de traitement du dossier de remise sur la TVA.	VAT Recovery Date - VAT Submission Date	Services Douanier
Release Lead Time	Durée nécessaire pour faire sortir une expédition d'un entrepôt douanier, son calcul dépend du régime douanier (FD ou VAT).	**Régime FD :** Release Date - Declaration Date **Régime VAT :** Release Date - VAT Recovery Date	CCA
Transport Lead Time	Durée nécessaire pour transporter l'expédition du port d'entrée vers la base opérationnelle	Reception Date - Delivery Date	CCA/Transporter
Restitution LeadTime	Durée nécessaire pour rendre le conteneur de la base opérationnelle vers le port d'entrée (calculé seulement dans le cas dès expédition de type "Sea Freight")	Restitution Date - Reception Date	CCA

TABLE 3.3 – Les différents lead times liés au processus d'importation

De plus, à travers les contrats établis avec les fournisseurs (les CCA), des objectifs ont été définis pour chaque lead time, ce point sera détaillé dans la partie

Ces lead times engendrent des coûts importants à la supply chain de l'entreprise, nous pouvons distinguer 5 coûts liés à l'importation d'une expédition : le coût de transport qui englobe le transport à l'international ainsi qu'au niveau national, le coût de stockage, la surestarie qui représente le coût appliqué aux conteneurs qui sont laissés au port durant la période de dédouanement, enfin nous trouvons les frais de dédouanement ainsi que les pénalités payées aux services de douanes dû aux retards ou des erreurs dans les dossiers présentés.

Chapitre 4 : Conception de la solution

Chapitre 4

Conception de la solution

Après avoir défini les outils nécessaires pour la résolution de notre problématique et détaillé le processus d'importation au sein de Schlumberger. Dans ce qui suit, nous allons présenter notre démarche de solution en appliquant les étapes du CRISP-DM.

Nous commencerons par la formulation du problème d'un point de vue Business dans la phase de compréhension des métiers, ensuite nous allons sélectionner les sources de données et les analyser afin d'avoir une idée sur les relations entre les colonnes. Nous allons par la suite entamer la partie de préparation des données pour les utiliser plus tard dans la modélisation par le Machine Learning. Enfin, nous allons déployer notre solution.

4.1 Compréhension des métiers

Avant d'entamer la partie technique de notre solution, nous devons d'abord aborder la problématique d'un point de vue business en découvrant les différents processus et les analyser.

4.1.1 Modélisation du processus d'importation

Afin de bien comprendre l'interaction entre les différentes parties prenantes du processus d'importation, nous avons opté pour la modélisation Business Process Model and Notation - BPMN 2.0. Notre problématique s'intéresse aux lead times au niveau national, alors lors de notre modélisation, on va s'intéresser aux différentes activités réalisées au niveau national.

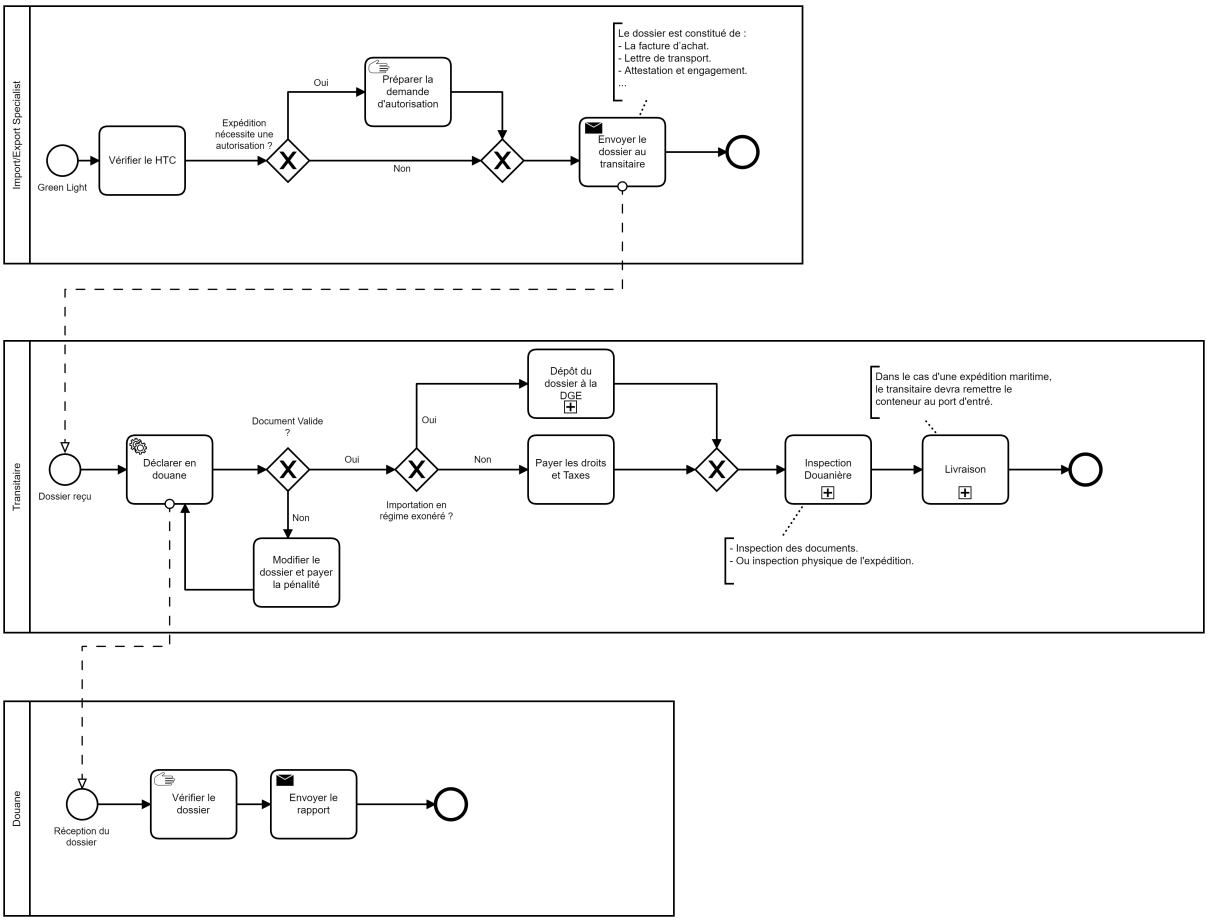


FIGURE 4.1 – Modélisation du processus d'importation au niveau de Schlumberger NAF.

La cartographie du processus nous a permis de voir de plus près les étapes du dédouanement ainsi que l'interaction entre les différents agents concernés. Afin d'analyser au mieux ce processus, nous allons développer un tableau de bord à l'aide de Power BI.

4.1.2 Analyse du processus

Afin de mieux comprendre le processus et suivre la performance de chaque partie prenante, nous avons opté pour la mise en place d'un tableau de bord Power BI. Nous avons tout d'abord défini le but de notre tableau de bord afin de déterminer les dimensions d'analyses et enfin les sources potentielles de données.
Le but du tableau de bord est d'analyser :

- Le nombre d'expéditions par moyen de transport et par transitaire.
- Les lead times en fonction de plusieurs paramètres dont le régime d'importation, la business line, le bureau de douane et par transitaire.
- Les coûts liés au stockage, dédouanement et les différentes pénalités.
- La performance des transitaires dans chaque phase de dédouanement.

Ces analyses seront agrégées en mois, trimestre et année. Nous avons ensuite mis en place une modélisation dimensionnelle en spécifiant les tables de dimensions et la table de faits :

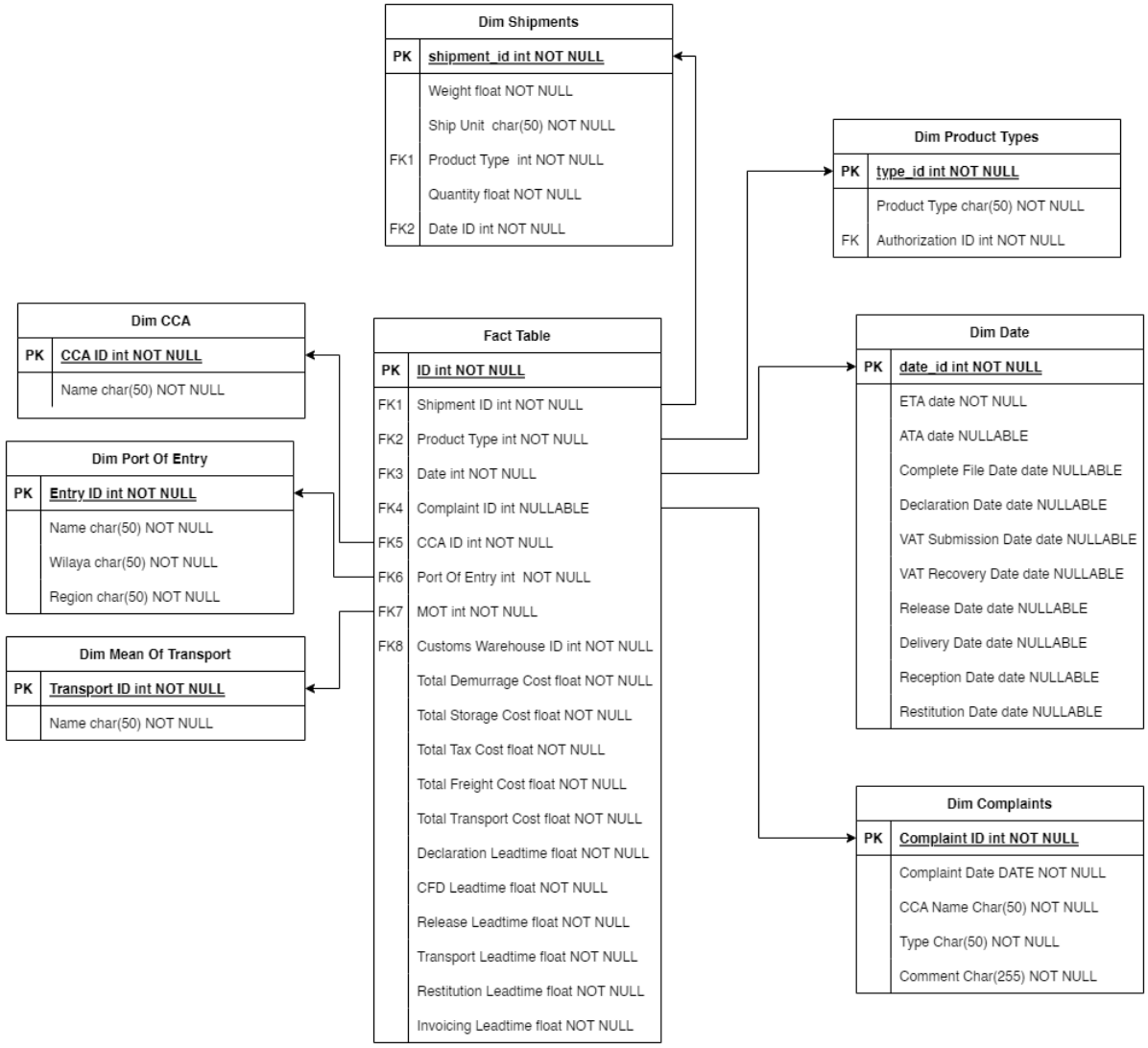


FIGURE 4.2 – Modélisation dimensionnelle proposée pour analyser le processus d'importation.

Passons maintenant aux sources de données disponibles :

- **L'ERP SAP :** SAP offre le meilleur niveau de détail possible avec une meilleure accessibilité aux expéditions, mais comme Schlumberger a migré vers SAP en mois d'avril, les données n'étaient pas d'une assez bonne qualité pour les utiliser.
- **Les rapports des transitaires :** ils n'englobaient pas les différentes dates importantes pour suivre la performance de la douane ou de l'équipe d'I/E de Schlumberger.
- **La clearance Portal :** C'est une plateforme interne à Schlumberger qui a été développé pour suivre l'évolution des expéditions depuis leur arrivée en Algérie. Elle est hébergée sur Microsoft Sharepoint qui est facilement intégrable avec Power BI.
- **Le Quest :** il est utilisé par Schlumberger pour signaler des réclamations sur la qualité du service.

Nous avons donc décidé d'utiliser la Clearance Portal pour analyser les expéditions et le Quest pour analyser les réclamations sur la qualité de service des transitaires.

Maintenant qu'on a déterminé nos sources de données, nous allons passer à Power BI afin

de développer le tableau de bord. Nous avons commencé par le traitement et la manipulation des données, ensuite, nous avons commencé à calculer les mesures nécessaires pour nos visualisations.

Le rapport Power BI est doté de plusieurs filtres par rapport au *transitaire*, *moyen de transport*, *régime d'importation*, et par *date d'arrivée de l'expédition*. Le rapport est constitué de quatre pages :

— Une page qui englobe le nombre d'expéditions dans chaque étape qui nous permettra de détecter des goulots, on a aussi ajouté le nombre d'expéditions par transitaire et par moyen de transport.

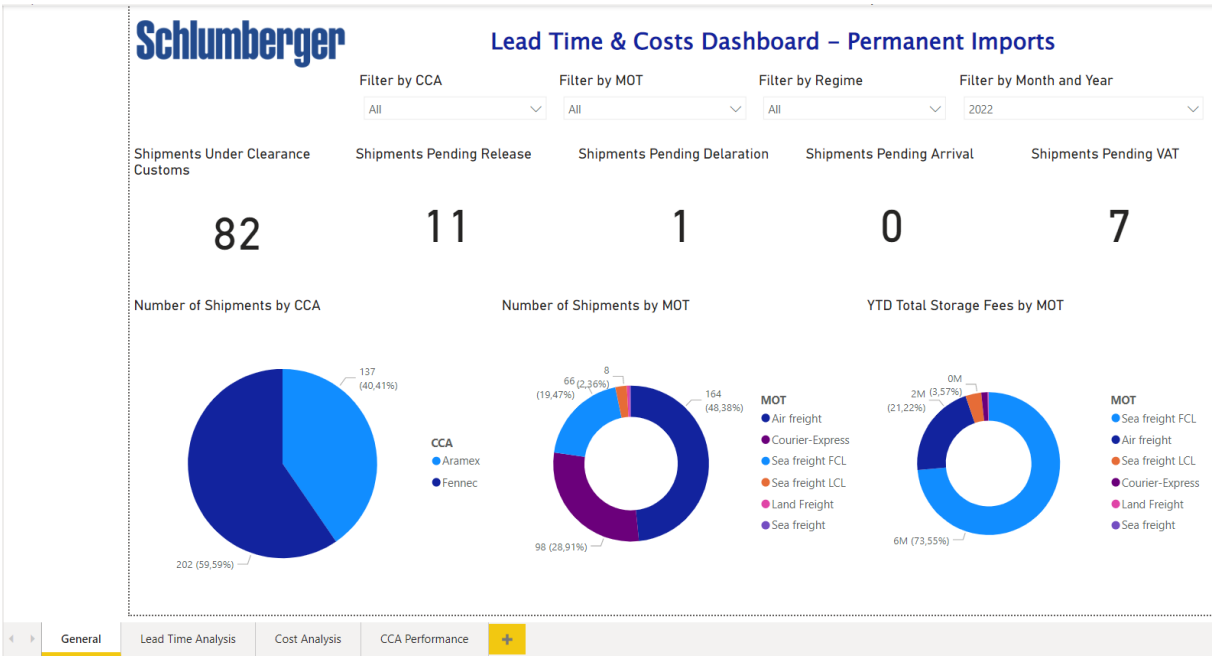


FIGURE 4.3 – Page D'accueil du rapport Power BI

Nous avons remarqué que le nouveau transitaire *Fennec* a géré 60% des expéditions de l'année 2022. Nous avons aussi constaté que le coût de stockage est le plus important (73% du coût de stockage total) dans le cas d'une expédition envoyé par voie maritime sachant que ce mode de transport ne représente que 20% des expéditions.

— Une page dédiée aux analyses des lead times par product line, par transitaire et par bureau de douane

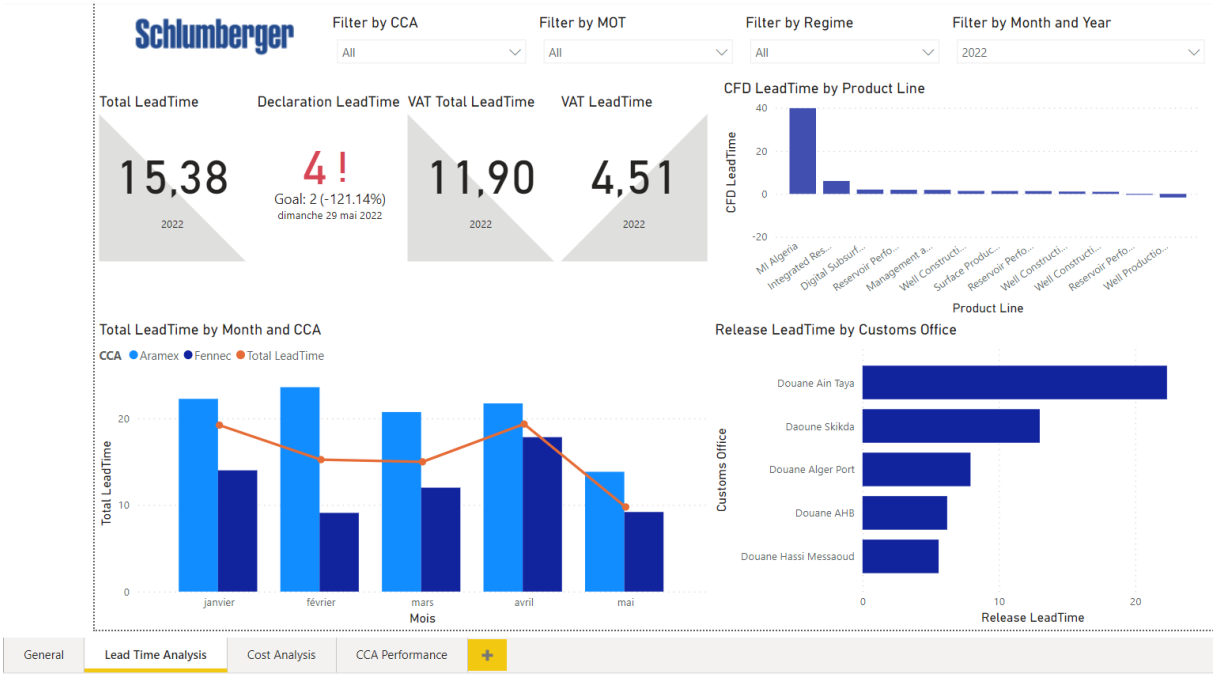


FIGURE 4.4 – Page d'analyse des lead times - rapport Power BI

Lors du développement du rapport Power BI, nous avons défini les objectifs en termes de lead times ciblés par Schlumberger, ce qui nous a permis de voir que les transitaires dépassent le lead time fixé de déclaration. Nous avons aussi remarqué que le temps moyen nécessaire pour préparer le dossier par les product lines de Schlumberger est en moyenne 2 jours, la product line *Well Production Systems* est très performante, car elle le prépare avant même l'arrivée de l'expédition, cependant, la product line *MI Algeria* est en sous performance, car elle prépare le dossier en 39 jours, après avoir analysé le problème de plus près en soulevant ce problème au niveau de Schlumberger, il s'est averé que cette product line importe les expéditions avec la remise documentaire et elle rencontre des problèmes avec la banque ce qui augmente les coûts de stockage et la surestarie.

— Une page dédiée aux analyses des différents coûts liés à l'importation.

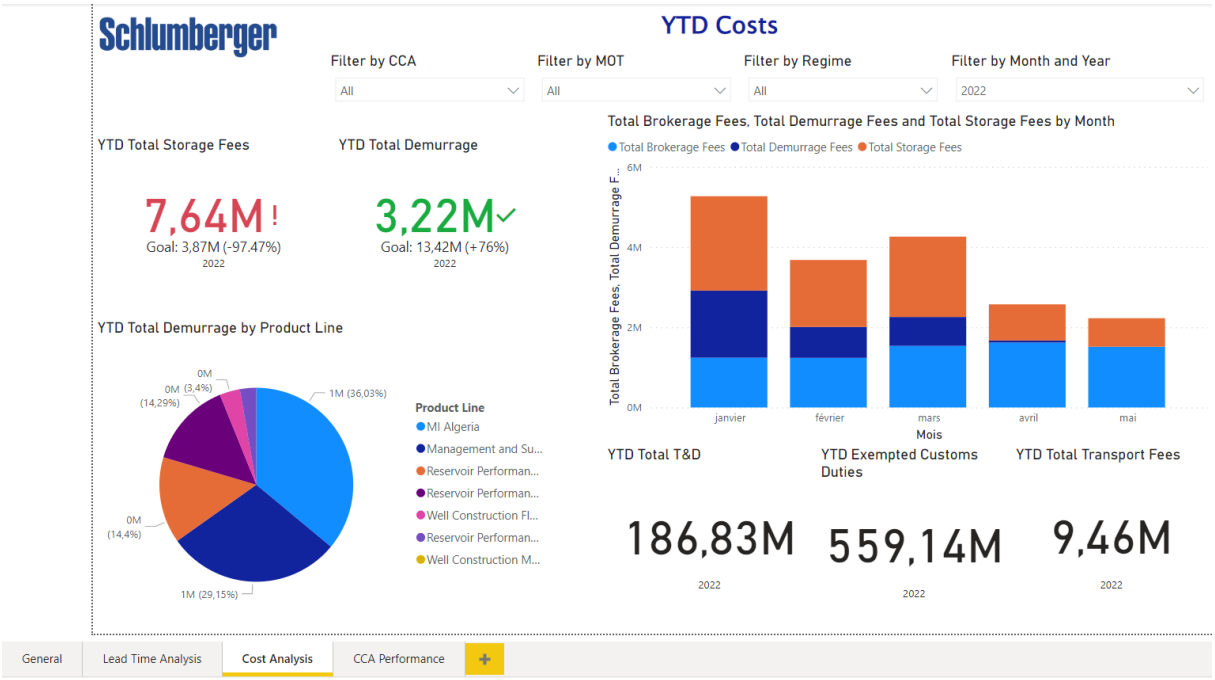


FIGURE 4.5 – Page d'analyse des coûts - rapport Power BI

Nous avons inclus les objectifs fixés par Schlumberger en termes de coût annuel de stockage et de surestarie pour l'année 2022 afin de pouvoir comparer avec les coûts réels. Nous avons remarqué que Schlumberger a déjà dépassé le coût de stockage ciblé pour l'année 2022, qui est dû au problème de la Product Line *MI Algeria*.

— Une page dédiée à la performance des transitaires :

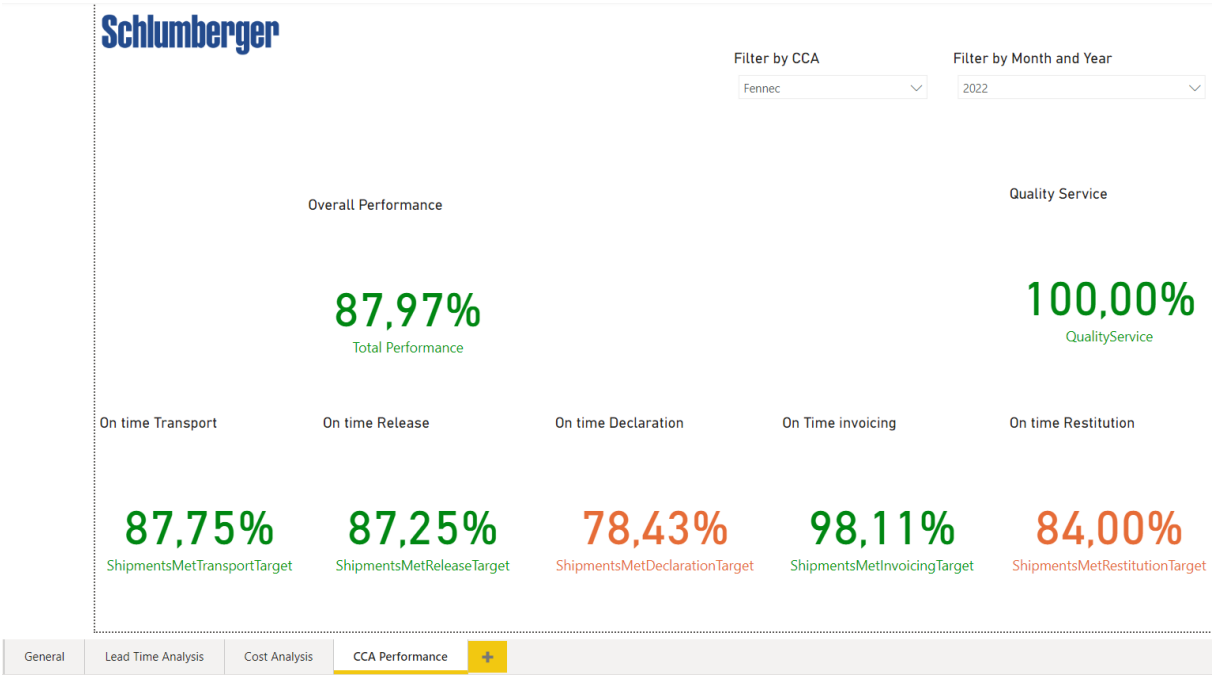


FIGURE 4.6 – Page d'analyse de la performance des transitaires - rapport Power BI

Nous avons pu calculer le score de chaque transitaire en se basant sur le scorecard qui a été définie dans les contrats entre les transitaires et Schlumberger (voir annexe

4.1.3 Analyse des transitaires : système de credit notes

Nous avons proposé d'implémenter un système de credit notes : c'est une écriture comptable à l'inverse de l'écriture initiale. Nous allons l'appliquer aux transitaires afin de les pénaliser.

Lors du calcul des credit notes, nous avons continué à utiliser les données de la Clearance Portal et nous avons choisi d'utiliser Python pour développer le script qui génère chaque mois les credit notes des transitaires et permet de calculer les pénalités.

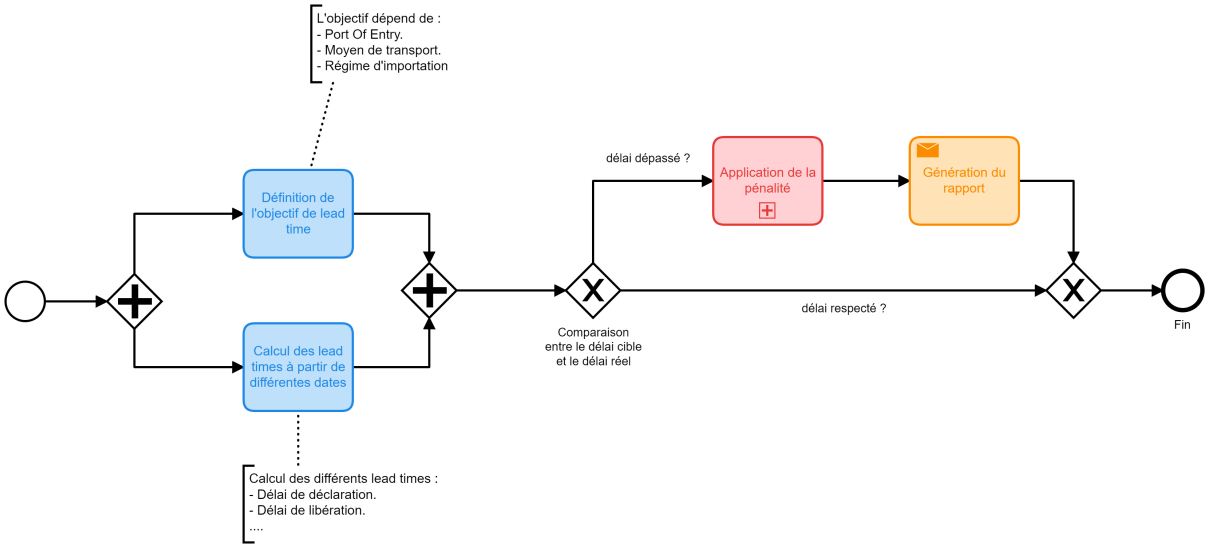


FIGURE 4.7 – Processus de génération du Credit Note des transitaires

L'application de la pénalité dépend de plusieurs autres paramètres, par exemple un retard d'une expédition qui arrive directement à Hassi Messaoud engendra une pénalité de 5% du coût total de la prestation du transitaire, cependant pour les autres entrées la pénalité est définie par le coût engendré pour chaque jour de retard.
Après avoir calculé les pénalités de chaque transitaire, nous allons ensuite générer deux rapports, un PDF pour afficher la pénalité total ainsi qu'un fichier Excel qui détaille la pénalité de chaque expédition. Le déploiement de cette solution sera détaillé plus tard dans la section déploiement de la solution.

Lors du diagnostic du processus d'importation, nous avons constaté que les Spécialiste I/Es au sein de Schlumberger reçoivent plusieurs requêtes qui demandent d'estimer les coûts et les lead times nécessaires pour importer un équipement ou une marchandise. Cette estimation paraît impossible pour les Spécialiste I/Es à cause du nombre de paramètres importants à prendre en compte dont la porte d'entrée, le moyen de transport, le type et le poids de l'expédition ainsi que le régime d'importation.

Afin de résoudre ce problème, nous avons proposé un modèle de Machine Learning pour estimer les lead times du processus d'importation en fonction des paramètres déjà cités. L'estimation des coûts se fera second temps en se basant sur les lead times estimés ainsi qu'aux caractéristiques intrinsèques de l'expédition.
En ce qui suit, nous allons continuer d'appliquer l'approche de CRISP-DM afin de

comprendre, nettoyer les données pour ensuite développer le modèle d'estimation et enfin le déployer.

4.2 Compréhension des données

Nous allons, à présent, explorer la source de données employée (Clearance Portal) et par la suite on effectuera un ensemble d'analyses statistiques sur ces données. L'étape de data understanding nous permettra de détecter les problèmes relatifs aux données afin de les nettoyer et les modéliser.

4.2.1 Collecte de données

Après avoir opté pour l'utilisation du "Clearance Portal" comme source de données vu qu'elle est plus complète et remplie fréquemment par les employés du département d'import-export de Schlumberger, nous avons décidé de ne garder que les données de l'année en cours (2022) car à partir de cette année, Schlumberger effectue ses opérations de dédouanement et de livraison des expéditions à travers deux transitaires : Aramex et Fennec, tandis qu'avant 2021, l'entreprise travaillait seulement avec Aramex, ce qui implique de très différentes performances.
De plus, après une première exploration des expéditions des années précédentes, nous avons remarqué que les données n'étaient pas homogènes et présentaient beaucoup de vide comparées aux données de l'année 2022, cela s'explique par le fait que l'intégration et l'usage de la Clearance Portal (la table sur MS Sharepoint) n'a commencé qu'en fin de 2021 au sein de l'entreprise (en Algérie), les données avant cette date ont été importées d'autres sources moins complètes et moins fiables, et donc ne seront pas considérées dans ce travail. Nous allons remédier à ce manque de données à travers une synthétisation (génération de données) efficace et relativement fiable qu'on présentera dans la suite de ce chapitre.
Les données exportées de MS Sharepoint contiennent les informations suivantes

— Des informations intrinsèques aux expéditions : le moyen de transport principal (MOT), le port d'entrée, le type de marchandise, le régime douanier, l'entreprise de transport, le transitaire choisi, le type de conteneur utilisé, la quantité, le poids, des informations sur la facturation, la base opérationnelle concerné, la business Line qui a initié la commande, l'entité légale, ainsi que des informations sur les différents coûts et frais de dédouanement (Frais de douane, surestarie, stockage, transport, etc).
— Les différentes dates relatives à l'importation présentées dans le chapitre 3 : "ATA", "CFD", "Declaration date", "Release date", "Delivery date" et "Reception date", pour les expéditions en régime d'exemption de TVA, nous avons en plus des dates précédentes, des dates relatives au traitement de la demande de remise sur la TVA par la douane.

En plus de ces informations, nous avons calculé les différents lead times définis dans le chapitre précédent pour chaque expédition, nous avons 4 nouvelles colonnes : “CFD lead time, “Declaration lead time”, “Release lead time" et “Transport lead time”. Ces colonnes représentent les lead times qu'on cherche à modéliser et à estimer dans les prochaines étapes.
Les données comportent 371 expéditions en 2022 (jusqu'au 15 mai 2022), cependant, nous allons nous intéresser seulement aux expéditions finalisées, c'est-à-dire celles qui ont un statut équivalent à “Shipment Closed” dans le cas d'expéditions finalisées (arrivées à une des bases opérationnelles de Hassi Messaoud) ou “Pending

Invoicing", qui indique qu'une expédition est arrivée à la base, mais la facturation n'a pas encore eu lieu. Les autres statuts ne nous intéressent pas pour le calcul des différents lead time. Ce filtrage est effectué à travers Python et la bibliothèque Pandas, on obtient après cette opération 263 expéditions.
Après avoir filtré les lignes, nous allons filtrer les colonnes : nous avons 92 colonnes (sans compter les colonnes des lead times crées), on utilisera qu'une partie de ces colonnes et on élimine les colonnes relatives aux identifiants, aux informations internes de Schlumberger ainsi que les colonnes relatives aux informations de facturation. Les colonnes utilisées seront présentées dans la suite du chapitre avec leur interprétation.

4.2.2 Analyse statistique des données

Après avoir récolté les données et n'avoir gardé que les données nécessaires pour ce travail, nous allons à présent effectuer une analyse statistique afin d'inférer des relations et des informations utiles sur nos données.
Pourcentage des données vides : La première étape est d'inspecter les vides dans notre jeu de données, la bibliothèque Pandas nous permet d'obtenir le pourcentage des vides dans chaque colonne sélectionnée. La figure 4.8 montre que certaines colonnes comme "MOT", "Status", "Regime", "Port of Entry", etc n'ont pas de vides, tandis que d'autres colonnes ont environ 17% de vides, les colonnes relatives aux coûts ont le plus de vides (plus de 70%). Les colonnes de dates (ATA, Declaration Date, Reception Date, etc) sont essentielles, elles contiennent environ 17% de valeurs non définies (vides), nous allons supprimer toute ligne ayant un vide dans une de ces dates, car il est impossible de calculer des lead times sans avoir toutes les dates pour chaque expédition.

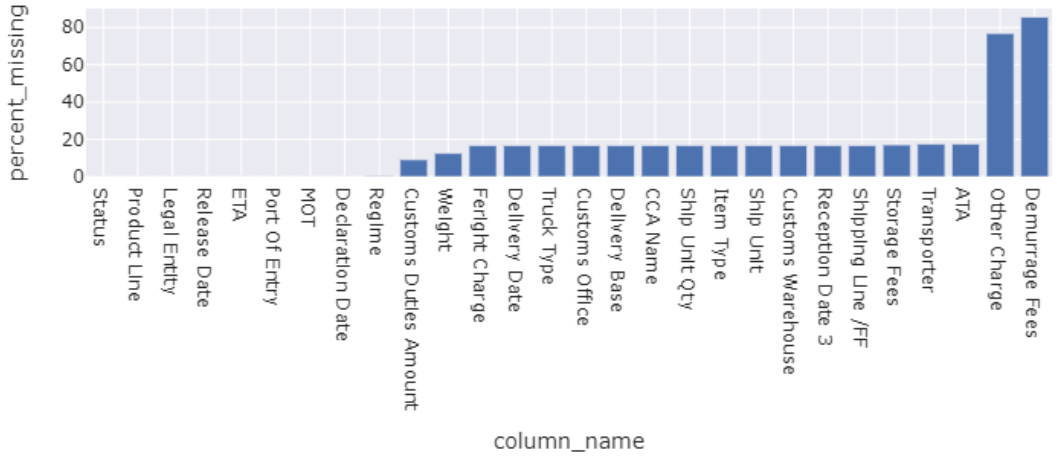


FIGURE 4.8 – Pourcentage (%) des valeurs non définies (vides) dans chaque colonne sélectionnée.

Distribution des données : Nous allons à présent visualiser la distribution empirique de certaines colonnes pour mieux comprendre la nature des expéditions, plus précisément : les colonnes "Product Line" avec "Legal Entity", "Port of Entry" avec "MOT" et "Regime", "Weight" avec "Ship Unit".

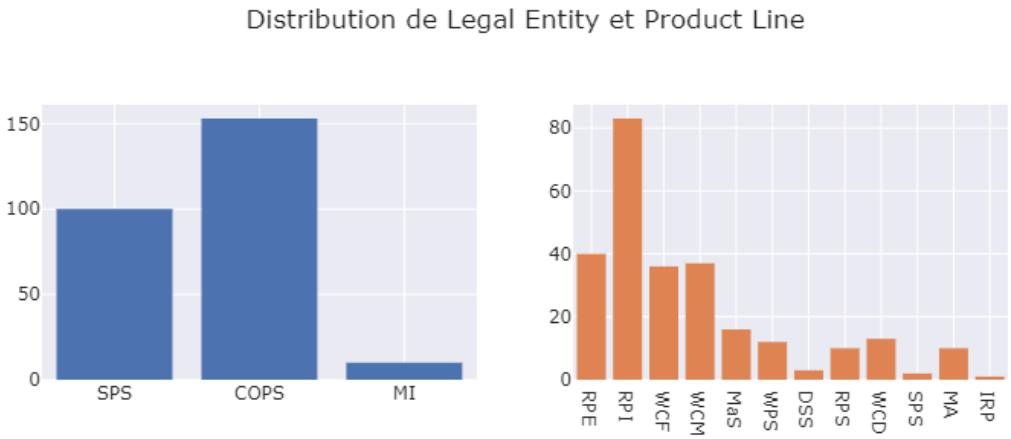


FIGURE 4.9 – Distribution de l'entité légale "Legal Entity" (à gauche) et de la Business Line "Product Line" (à droite)

On remarque à partir de la figure 4.9 que 83 expéditions sont commandées par la Business Line RPI (Reservoir Performance Intervention) tandis que la Business Line IRP (Integrated Reservoir Performance) comporte une seule expédition seulement. On trouve aussi que les deux entités légales commandent la majorité des expéditions tandis que MI-Algérie représente moins de 5% des expéditions, cette entité légale ne va pas donc influencer sur les estimations et nous pouvons l'ignorer.

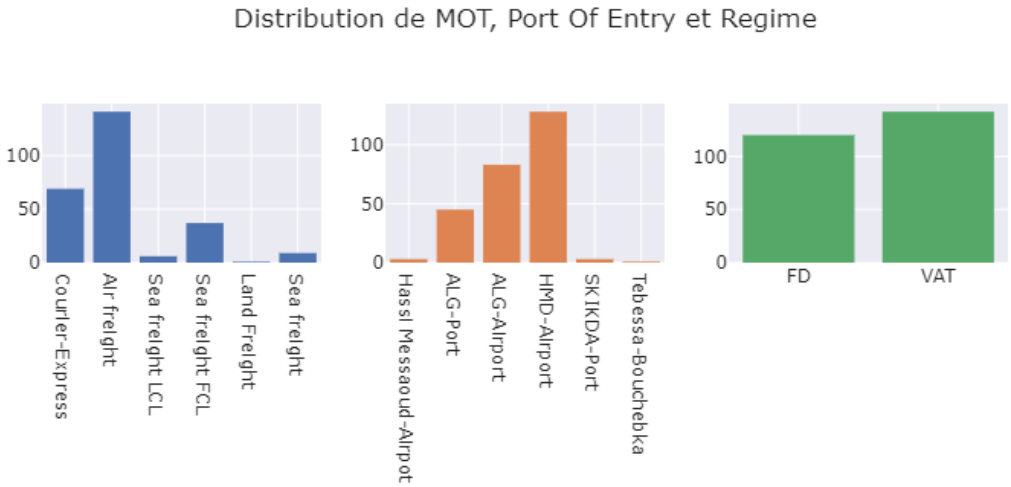


FIGURE 4.10 – Distribution du moyen de transport "MOT" (à gauche), du port d'entrée "Port of Entry" (centre) et du régime douanier "Regime" (à droite)

Nous remarquons aussi à travers la figure 4.10 qu'une bonne partie (presque 50%) des expéditions sont transportées par fret aérien avec l'aéroport de Hassi Messaoud comme port d'entrée commun, on retrouve aussi environ 16% des expéditions sont transportées par fret maritime, principalement en FCL (Full Container Load). Pour le régime douanier, on retrouve que les deux régimes "VAT" et "FD" sont répartis équitablement (53% et 46% respectivement).

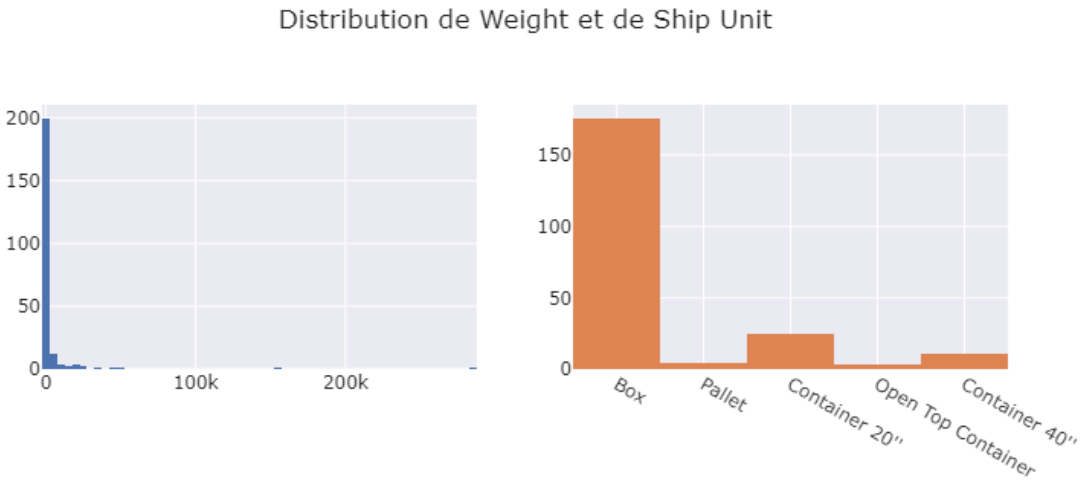


FIGURE 4.11 – Distribution du poids "Weight" (à gauche) et de l'unité d'expédition "Ship Unit" (à droite)

Selon la figure 4.11, on remarque que la majorité des expéditions ont un poids inférieur à 2000 Kg, ceci est démontré aussi par l'unité d'expédition "Ship Unit" où on remarque que l'unité "Box" ou boite représente aussi 176 expéditions (environ 67% de toutes les expéditions) car la majorité des expéditions ont un poids relativement petit.

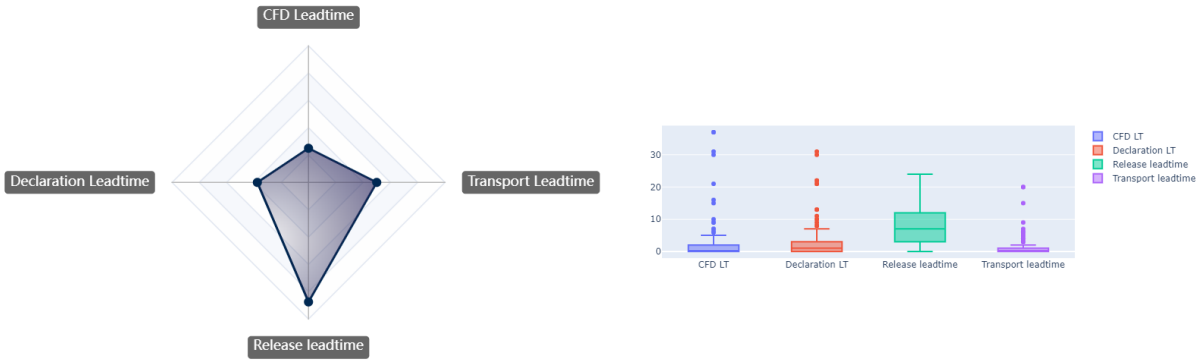


FIGURE 4.12 – Analyse de la distribution et de la proportion des lead times d'importation

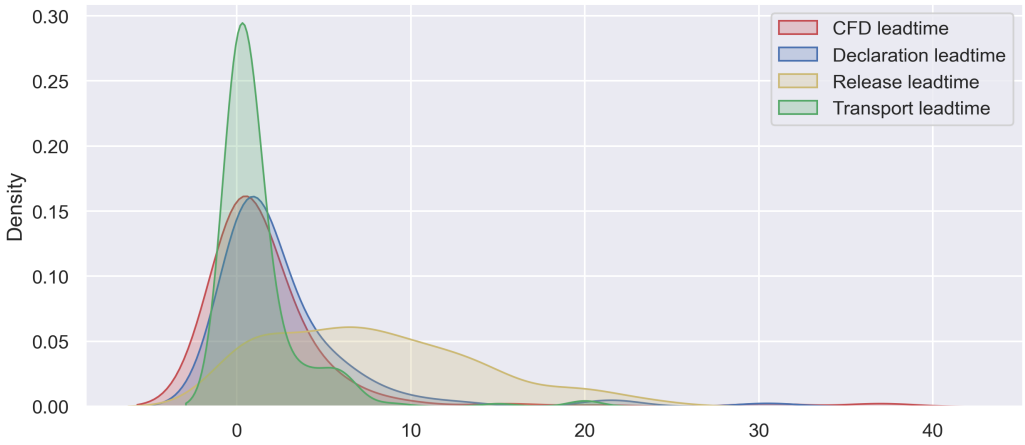


FIGURE 4.13 – Distribution estimée des différents lead times

En ce qui concerne les lead times, la figure 4.12 présente la moyenne de chaque lead time, on remarque que le "Release lead time" est le plus lent, avec une moyenne de 7.6 jours comparés aux autres lead times, on conclut alors que cette étape relative au traitement du dossier par la douane est une source importante de retard tout au long du processus d'importation et il est important de l'estimer. On remarque aussi selon la boite à moustache dans la même figure que le CFD lead time comporte quelques valeurs aberrantes arrivant jusqu'à 31 jours.
Pour finir, la figure 4.13 montre la distribution estimée (kernel density) des différents lead times, on remarque que le lead time "Release lead time" ne suit pas une loi normale, tandis que les autres lead times suivent une loi normale, mais avec une asymétrie droite qui peut être négligé par rapport à sa concentration.
Analyse des effets : Il est utile d'étudier l'effet entre les différents lead times et les autres colonnes (facteurs) afin de déterminer quelles sont les variables potentiels qui influencent sur ces lead times. La majorité des variables sont catégoriques (nominale ou ordinale) tandis que les variables dépendantes sont numériques (les lead times). Afin de mesurer cet effet dans cette situation, on utilise une analyse de type ANOVA, cependant, l'ANOVA classique nécessite un ensemble de conditions à vérifier dans les données :

— **Normalité** : La figure 4.13 indique que la distribution des lead times est normale (même si elle est asymétrique, sa concentration reste élevée), de plus, la taille de l'échantillon de 186 exemplaires renforce l'hypothèse de normalité, nous pouvons donc conclure que la distribution est normale avec une asymétrie qui peut être négligée.
— **Homoscédasticité :** l'ANOVA classique nécessite que la variance entre les groupes soit égales, pour cela, nous avons effectué le test de Levene comme le montre le tableau 4.1 entre les lead times et un ensemble de colonnes, on remarque que certains groupes ont une variance égale, mais ce n'est pas le cas pour toutes les colonnes, par exemple la p-valeur de "CCA Name" avec "Declaration lead time" est inférieur à 0.05, dans ce cas, cette colonne est dotée d'une hétéroscédasticité, c'est le cas aussi pour d'autres colonnes, la condition d'homoscédasticité n'est donc pas vérifiée, l'annexe A présente le test de Levene avec plus de détails.

Lead time \ Colonne	CCA Name	MOT	Regime
Declaration lead time	0.000143	0.212268	0.982939
Release lead time	0.311956	0.040687	0.073752
Transport lead time	0.049220	0.148590	0.940149

TABLE 4.1 – Test de Levene sur les lead times avec "CCA Name", "MOT", "Regime"

La condition d'homoscédasticité des groupes de l'ANOVA n'est pas vérifiée, nous n'avons pas besoin de vérifier les autres conditions et il n'est pas possible d'appliquer l'ANOVA classique (voir l'annexe A pour plus de détails), c'est pour cela, qu'on a employé une variante de cette méthode : ANOVA de Kruskal-Wallis, elle permet de déterminer les variables qui influencent sur la variabilité des lead times sans prendre en considération les conditions précédentes (cette méthode est détaillée dans l'annexe A). La table 4.2 nous donne une idée sur les variables qui influencent le plus sûr les lead times (à travers le calcule de η^2 qui se base sur la statistique H calculée par le test et qui représente le niveau d'influence d'une dimension).

(a) Pour CFD lead time

Source	η^2
Product Line	0.111807
Legal Entity	0.091130
Customs Office	0.052979
Port Of Entry	0.029962
Shipping Line /FF	0.029704

(b) Pour Declaration lead time

Source	η^2
CCA Name	0.224742
Transporter	0.187482
Shipping Line /FF	0.102934
Legal Entity	0.066495
MOT	0.062672

(c) Pour Release lead time

Source	η^2
Regime	0.281433
Transporter	0.267920
Shipping Line /FF	0.199596
CCA Name	0.096268
MOT	0.094842

(d) Pour Transport lead time

Source	η^2
MOT	0.106028
Port Of Entry	0.078817
Product Line	0.071038
Shipping Line /FF	0.063303
Transporter	0.055862

TABLE 4.2 – Les 5 dimensions (colonnes) avec le plus grand effet sur chaque lead time selon le test de Kruskal-Wallis

4.3 Préparation des données

Maintenant que nous avons analysé les données, nous allons entamer la partie de traitement des données afin de pouvoir les charger dans les différents modèles de Machine Learning.
Lors de la préparation des données, nous avons effectué deux grandes tâches : le nettoyage des données et la synthétisation des données (génération des données) pour augmenter leur volume.

4.3.1 Data Cleaning (Nettoyage des données)

Cette partie consiste à transformer et nettoyer les données exportées depuis la Clearance Portal pour les utiliser dans les différents modèles de Machine Learning. Après la réalisation d'une analyse approfondie des données (section 2.2), nous avons constaté quelques problèmes dans ces derniers :

- Les données de 2021 ne sont pas fiables, car la validation de données n'était pas encore implémentée dans la Clearance Portal, par exemple, on retrouve expéditions avec une date de réception à Hassi Messaoud avant même la date d'arrivée de l'expédition en Algérie.
- Il existe plusieurs expéditions qui sont toujours en cours de dédouanement, donc plusieurs dates sont vides.
- Il existe des colonnes qui ont un format de date différent (à la place du format *JJ/MM/AAAA*, on trouve le format *MM/JJ/AAAA*).

Nous avons supprimé les expéditions de 2021 ainsi que les expéditions qui sont en cours de dédouanement et nous avons transformé les dates ayant un format différent des autres colonnes pour pouvoir calculer les différents lead times présenté par les colonnes : "CFD lead time", "Declaration lead time","Release lead time", "Transport lead time".

Maintenant que nos données ne présentent plus de vides, nous avons calculé les différents lead times en calculant la différence en jours entre les dates et en éliminant bien sûr les jours chômés (les week-ends) car les objectifs sont définis en se basant sur les jours ouvrés.

Nous allons à présent entamer la préparation des données pour les charger dans les différents modèles de Machine Learning. La préparation se fait en deux parties :

1. **Sélection des colonnes :** nous devons choisir les colonnes à faire passer à notre modèle, les dimensions sélectionnées sont :

Dimension	Description
Status	Le statut de l'expédition
Product Line	La Business unit qui a demandé l'expédition
Legal Entity	L'entité légale qui a demandé l'expédition
Weight	Le poids de l'expédition en kg
Ship Unit Qty	La quantité demandée
Ship Unit	L'unité logistique (conteneur, palette, etc.)
Type du produit	Le type de l'expédition (chimique, explosif, etc.)
Port Of Entry	La région d'entrée de l'expédition (Alger, Skikda, etc.)
MOT	Le moyen de transport
Regime	Le régime d'importation (exonéré ou pas)
Service Level	L'urgence de l'expédition (urgent ou pas)
Delivery Base	La destination finale de l'expédition (HMD ou autre)
Customs Office	Le bureau responsable de dédouanement
Customs Warehouse	L'entrepôt douanier
Transitaire	Le nom du transitaire (Aramex ou Fennec)
Le transporteur	Le nom du transporteur de l'expédition en Algérie

TABLE 4.3 – Les dimensions utilisées pour les modèles de Machine Learning

2. **Création des données d'apprentissage et les données d'évaluation :** afin d'éviter le sur-apprentissage des modèles de Machine Learning, nous avons opté

pour une cross validation avec 4 folds. (Voir Annexe B pour plus de détails).

4.3.2 Génération des données

Nous avons précédemment remarqué que nous n'avons que 180 expéditions utilisables pour la modélisation, ce faible volume de données peut causer deux problèmes :

— **Sur-apprentissage (*Overfitting*) :** Le modèle peut tomber dans une situation de sur-apprentissage (overfitting), il ne pourra pas apprendre la tendance générale des données et va s'adapter à chaque cas dans les données fournies. Cette situation est très fréquente en Machine Learning et produit des modèles instables avec des performances qui varient significativement entre différentes partitions de données durant la validation croisée (cross validation). En ajoutant plus de données, cette situation pourra être évitée ou du moins minimisée. On verra par la suite que l'entraînement d'un modèle linéaire sur des jeux de données de taille plus importante réduit l'écart type de l'erreur et permet d'obtenir des modèles plus stables.

— **Faible précision :** Un faible volume de données ne permet pas au modèle d'apprendre les relations et la tendance des données, et donc ne permet pas de réduire l'erreur suffisamment.

Cependant, bien qu'en théorie, plus de données permet d'éviter ces deux problèmes, il est dur de réaliser cela en pratique : ajouter des données peut aussi rajouter du bruit et rendre le modèle moins précis, ou encore le biaiser si les données ajoutées ne sont pas bien distribuées. C'est pour cela que qu'il est nécessaire de rajouter des données de qualité, or, des données qui suivent une distribution similaire aux données initiales tout en étant diversifiées.

Afin d'assurer une bonne qualité des données, nous allons générer (synthétiser) des données de deux manières : à travers une simulation à événements discrets et à travers du deep learning en employant un réseau de neurones de type TVAE (Tabular Variational Auto-Encoder). Pour chaque méthode, nous évaluerons la similarité des données générées avec les données réelles et nous les fusionnerons pour créer 3 jeux de données différents : les données initiales seulement D_1, les données initiales fusionnées avec les données générées par la simulation D_2 et les données initiales fusionnées avec les données générées par le réseau de neurones TVAE D_3. Ces différents jeux de données seront utilisés pour la modélisation dans la prochaine phase.

À travers la simulation

En utilisant le logiciel AnyLogic (voir l'annexe

D'abord, nous avons ajouté une carte GIS centrée sur l'Algérie dans notre environnement, cette carte GIS nous permettra de placer les différents agents (port d'alger, base opérationnelle, etc) et avoir des visualisations lors de la simulation comme le montre la figure 4.14.

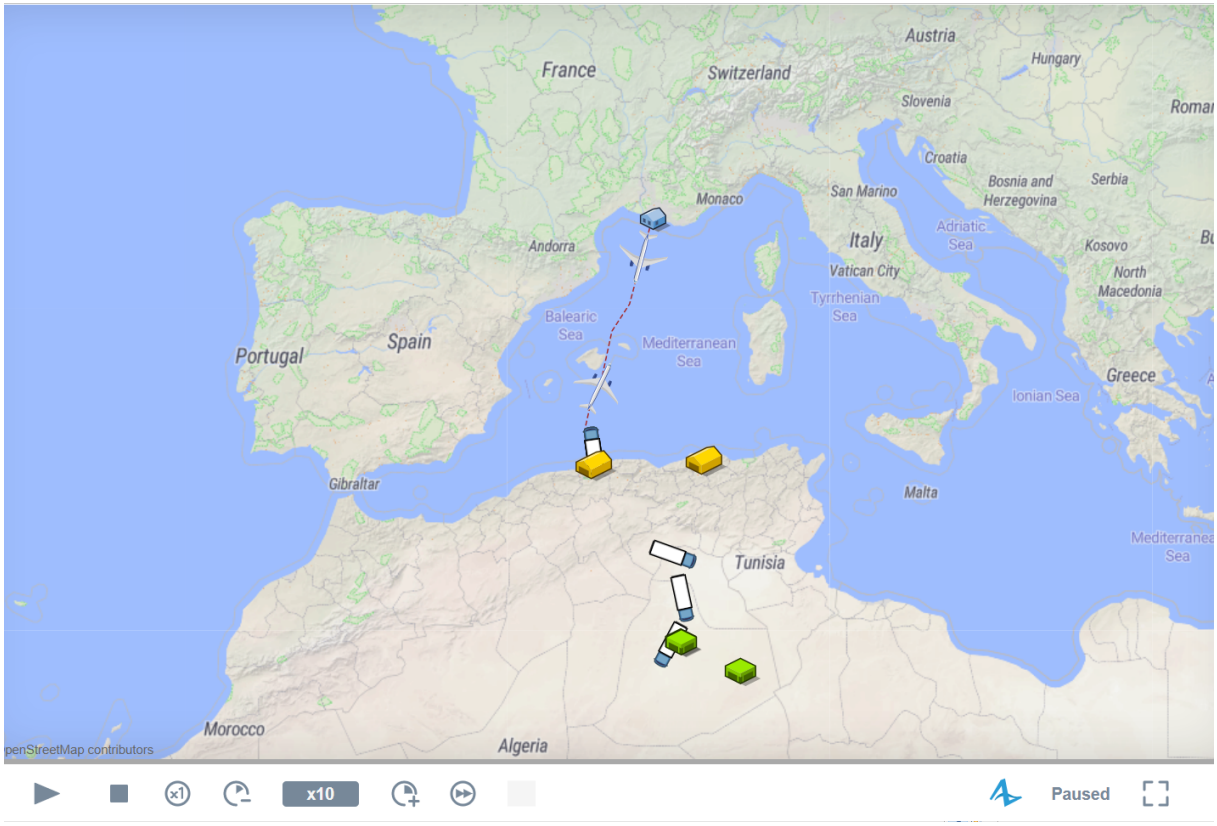


FIGURE 4.14 – Carte GIS sur l'environnement de simulation.

Ensuite, nous avons créé plusieurs agents : un agent pour représenter la douane du port d'alger et un autre pour le port de Skikda, d'autres agents pour représenter les bases opérationnelles, et un groupe d'agents pour représenter les véhicules et les avions. Dans cette simulation, nous avons pris en considération que le port d'Alger et de Skikda car la version de AnyLogic utilisée est une version gratuite limitée, mais il est possible de rajouter plus d'agents dans la version professionnelle. La figure 4.15 permet d'avoir une vue globale sur l'ensemble des agents (et des autres objets créés) dans l'environnement.

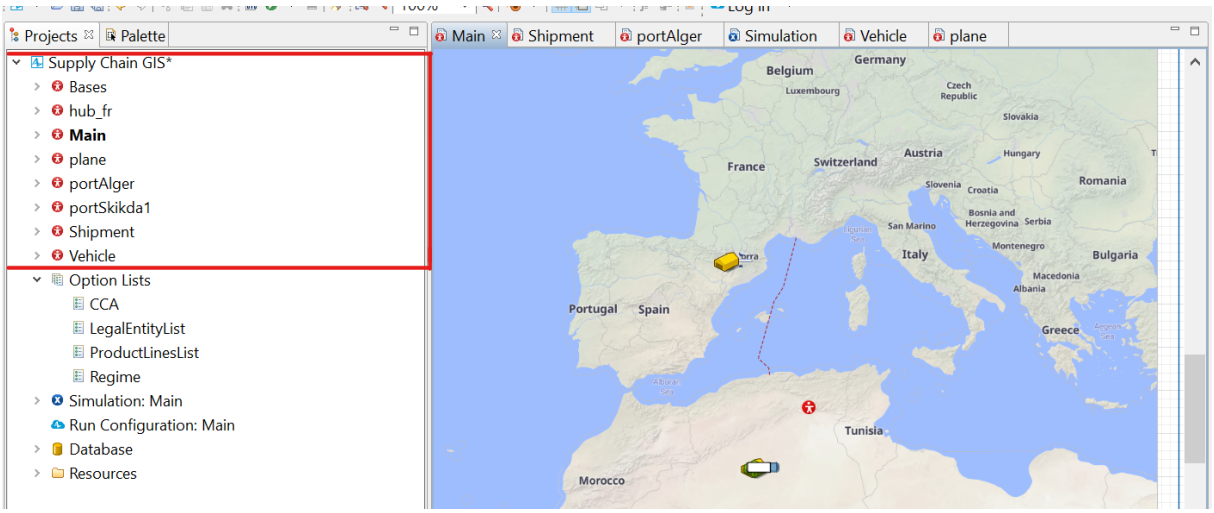


FIGURE 4.15 – Interface générale de AnyLogic et le panneau des objets de la simulation.

Pour le délai (lead time) de chaque étape, nous avons utilisé les distributions récupérées à partir des données initiales par rapport aux lead times : "CFD lead

time", "Declaration lead time", "Release lead time", "Transport lead time". Nous avons utilisé la fonction "CustomDistribution" de AnyLogic pour implémenter une fonction en Java pour calculer ces délais à partir de ces distributions empiriques, cette procédure nous permet d'obtenir des délais similaires aux délais réels, la figure 4.16 montre un exemple de ces méthodes : le fonctionnement de la méthode qui calcule le délai que prend un transitaire pour déclarer une expédition au niveau de la douane "Declaration lead time".

```
public double function get_declaration_lt(Shipment shipment)
{
        double delay;

        int[] leadtimes;
        switch (shipment.cca) {
                case Mouldi_Meddeb :
                    leadtimes = selectFrom( cca_mm_dist_declaration ).arrayOfInt(cca_mm_dist_declaration.obs);
                    break;
                case Fennec:
                    leadtimes = selectFrom( cca_fennec_dist_declaration ).arrayOfInt(cca_fennec_dist_declaration.obs);
                    break;
                default:
                    return 2;
            }

        CustomDistribution dist = new CustomDistribution(leadtimes);
        delay = dist.get(new Random(0));
        traceln(delay);
        return delay;
}
```

FIGURE 4.16 – Méthode en Java qui permet de simuler le lead time de déclaration à la douane en suivant la distribution des données initiales.

Nous avons développé une méthode pour chaque étape (chaque bloc "Service") pour simuler tous les lead times, nous avons aussi ajouté des attributs tels que le poids "Weight", la Business Line "Product Line", l'entité légale "Legal Entity" et les autres colonnes à l'agent de type "Shipment" qui représente les expéditions, ces attributs sont affectés à partir des distributions des données réelles, tout en prenant en considérations les différentes contraintes entre chaque attribut :

1. Les expéditions avec des poids importants ont une unité d'expédition de type "Container 20"" ou "Container 40"" et sont livrées par fret maritime.
2. Les expéditions livrées par fret maritime ont un port d'entrée qui correspond au port d'Alger ou au port de Skikda, celles livrées par le fret aérien ont "Aéroport HMD" ou "Aéroport AHB" comme port commun, etc.
3. Les attributs représentant l'entrepôt douanier et le bureau de douane sont aussi affectés selon le port d'entrée déterminé par la contrainte précédente.
4. Le transporteur est affecté selon le transitaire attribué aléatoirement par la simulation.

Après avoir fini cette modélisation, on lance la simulation sur AnyLogic et on augmente la vitesse de simulation, au bout de 10 minutes, on obtient environ 700 expéditions simulées (figure 4.17) et on exporte les résultats enregistrés dans la base de données de AnyLogic vers un fichier Excel pour les utiliser dans la modélisation.

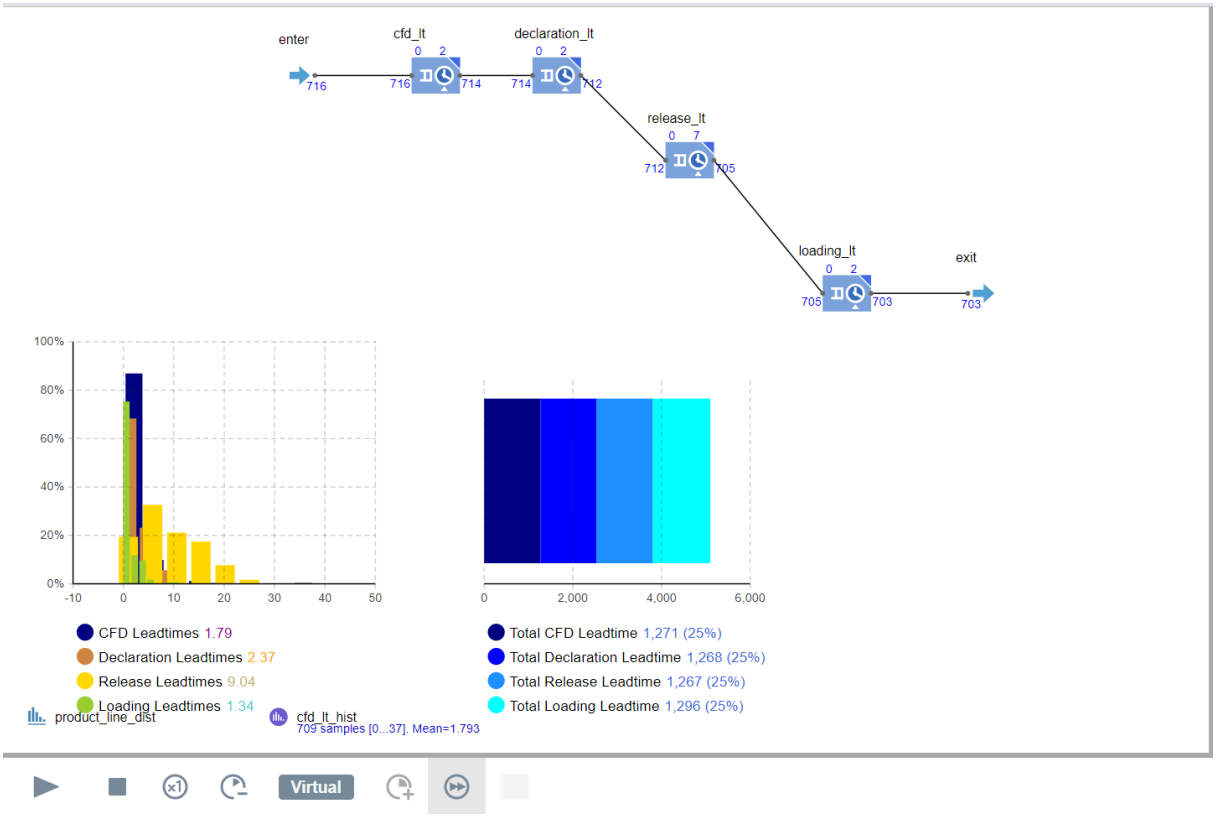


FIGURE 4.17 – Résultat de la simulation du processus de dédouanement avec ses différentes étapes.

Génération à travers les réseaux de neurones TVAE

La deuxième méthode de synthétisation des données employée se base sur les réseaux de neurones de type “Tabular Variational Auto Encoders” (voir C pour plus de détails sur leurs fonctionnement), ces réseaux peuvent prendre en entrée un jeu de données tabulaire et apprendre les différentes distributions de chaque colonne ainsi que les corrélations entre ces colonnes. Nous avons utilisé Python ainsi que la bibliothèque *sdv* (*Synthetic Data Vault*) qui contient un réseau TVAE qui permet de générer des données facilement. Nous avons aussi développé une fonction en Python pour prendre en considération les différentes contraintes logiques entre chaque colonne comme mentionné précédemment dans la partie de la simulation (par exemple les contraintes entre le poids et l’unité d’expédition, le moyen de transport et le bureau de douane, etc).

```python
from sdv.tabular import TVAE
from sdv.constraints import FixedCombinations
from sdv.constraints import Positive
from sdv.constraints import Between

mot_poe_constraint = FixedCombinations(
    column_names=['MOT', 'Port Of Entry','Delivery Base','Customs Warehouse','Customs Office','Weight','Ship Unit','Ship Unit Qty

full_delay_constraint = Between(column='Full delay',low=1,high=50,handling_strategy='transform')

truck_constraint = FixedCombinations(
    column_names=['CCA Name','Transporter','Transport leadtime'],handling_strategy='transform')

release_constraint = FixedCombinations(
    column_names=['Regime','Release leadtime'],handling_strategy='transform')

declaration_constraint = FixedCombinations(
    column_names=['Legal Entity','Product Line','CFD leadtime'],handling_strategy='transform')

positive_constraint = Positive(columns=['CFD leadtime','Declaration leadtime','Release leadtime',
        'Loading leadtime','Transport leadtime'],strict=False,handling_strategy='reject_sampling')

constraints = [mot_poe_constraint,
               #weight_constraint,
               truck_constraint,
               release_constraint,
               declaration_constraint,
              positive_constraint]

model = TVAE()
model.sample_conditions = constraints
model.fit(data)
```

FIGURE 4.18 – Implémentation du réseau générative TVAE avec les contraintes logiques.

En sortie du réseau, on obtient environ 500 nouvelles lignes de données avec toutes les colonnes générées (les colonnes des lead times aussi).
Cette méthode à deux avantages importants comparée à la génération de données à travers la simulation :

1. L'application de cette méthode est facile et ne prend pas de temps, tandis que pour la première méthode, il a fallu simuler tout le processus manuellement.
2. Les TVAE peuvent apprendre les différentes corrélations entre les variables, ce qui est impossible à faire automatiquement à travers la simulation.

Il existe d'autres types de réseaux modernes tels que le CTGAN, mais le TVAE est le réseau le plus adapté à nos besoins, car il est facile à implémenter, à comprendre et ses résultats sont satisfaisants comme on le verra par la suite.

Évaluation des données générées

Après avoir généré des données, on évalue leur qualité, pour cela, la bibliothèque SDV offre une méthode *"evaluate"* qui prend en paramètre les données réelles et les données générées et évalue leur similarité à travers plusieurs mesures : des mesures statistiques (divergence de Kullback-Leibler, vraisemblance, test de Kolmogorov-Smirnov, etc), des modèles linéaires, et d'autres méthodes. Cependant, nous utilisons que la divergence de Kullback-Leibler D_{KL} (voir C pour plus de détails sur le formalisme mathématique de cette mesure) afin de simplifier le calcule, le tableau 4.4 présente les scores de similarité à travers cette méthode :

Méthode	Similarité
Simulation	45.35%
TVAE	80.94%

TABLE 4.4 – Scores de similarité entre les données initiales et les données générées avec les deux méthodes à travers la divergence de Kullback-Leibler D_{KL}

On remarque que les données générées avec la méthode du TVAE est plus proche de la réalité, ceci est principalement due aux raisons citées précédemment, le réseau TVAE peut apprendre les corrélations entre les colonnes et des distributions multidimensionnelles, tandis que la simulation ne peut pas les représenter.
Après avoir nettoyé les données et généré un volume plus important, nous avons à présent 3 jeux de données : D_1 qui comporte que les données originales avec

187 expéditions, D_2 représentant les données originales fusionnées avec les données simulées avec 687 expéditions, D_3 représentant les données originales fusionnées avec les données générées par le TVAE avec 687 expéditions prêt à être utilisés pour la modélisation.

4.4 Modélisation

Dans cette partie, nous allons développer et évaluer des modèles de régression pour estimer les 4 lead times : y_{cfd} pour le "CFD lead time", y_{dec} pour le "Declaration lead time", y_{rel} pour le "Release lead time", y_{transp} pour le "Transport lead time". On entraînera chaque modèle sur 3 jeux de données différents :

1. D_1 : Représente les données originales uniquement.
2. D_2 : Représente les données originales fusionnées avec les données générées à travers la simulation.
3. D_3 : Représente les données originales fusionnées avec les données générées à partir du réseau de neurones TVAE.

4.4.1 Estimation des lead times

Modèle de base de référence et métrique d'erreur

Afin de déterminer si un modèle est performant ou pas, nous allons définir un modèle basique de référence, utilisé *mentalement* par les employés de Schlumberger pour estimer les lead times, ce modèle consiste tout simplement à estimer les lead times par la moyenne des lead times précédent. Pour chaque lead time, nous allons présenter les performances de ces modèles sur les 3 jeux de données D_1, D_2, D_3 , la performance sera mesurée à travers l'erreur moyenne absolue MAE :

$$MAE = \frac{1}{n}\sum_{i=0}^{n} |\hat{y}_i - y_i|$$

Pour l'erreur de référence, il suffit de remplacer $\hat{y}_i$ par la moyenne $\bar{y}$ on obtient donc :

$$MAE_{\text{réf}} = \frac{1}{n}\sum_{i=0}^{n} |\bar{y}_i - y_i|$$

Cette erreur est facile à interpréter, car elle est dotée de la même unité des lead times (jours). Par la suite, on comparera l'erreur de référence du modèle basique avec l'erreur de chaque modèle et on prendra le modèle qui présente l'erreur MAE la plus inférieure à cette erreur de référence.

Estimation du CFD lead time

Nous allons commencer par l'entraînement du modèle d'estimation du Complete File Date lead time, il est à noter que l'équipe de Schlumberger est la seule responsable de la préparation du dossier de l'expédition, il est donc inutile d'intégrer les colonnes qui sont externes à Schlumberger comme le transitaire, le bureau de douane et autres. Les colonnes qui seront utilisées pour l'estimation de ce lead time sont : *Product Line, Legal Entity, Ship Unit Qty, Ship Unit, Item Type, MOT, Service level.*
Après avoir entraîné les modèles de Machine Learning sur nos jeux de données, nous avons trouvé les résultats présentés dans le tableau 4.5

Modèle	D_1	D_2	D_3
Support Vector Machine	1.94	0.747	0.8492
Linear Regression	1.71	0.89	0.96
Random Forest	1.92	0.91	1.03
Gradient Boosted Trees	1.86	0.92	1.04
Decision Tree	1.95	1.00	1.10
Modèle de Base	2.47	2.47	2.47

TABLE 4.5 – MAE des modèles entraînés et évalués sur 3 jeux de données pour "CFD lead time"

Le meilleur modèle d'estimation du CFD lead time est le modèle *Support Vector Machine* avec une erreur MAE de 0.74 jour qui est l'équivalent de 17 heures. Le modèle le plus performant a été trouvé avec le jeu de données générées à l'aide de la simulation.

Estimation du Declaration lead time

Passons maintenant à l'estimation du Declaration lead time qui représente le temps que prend les transitaires pour déposer le dossier au bureau de douane. Les colonnes qui seront utilisées lors de l'estimation de ce lead time sont : *Ship Unit, Item Type, MOT, Regime, Service level, Customs Office, CCA Name, Declaration lead time*
Après avoir entrainé les mêmes modèles utilisés lors de l'estimation du CFD lead time, on trouve les résultats présentés dans le tableau suivant :

Modèle	D_1	D_2	D_3
Support Vector Machine	2.00	1.12	1.04
Linear Regression	2.43	1.22	1.16
Random Forest	2.28	1.25	1.19
Gradient Boosted Trees	2.32	1.26	1.20
Decision Tree	2.30	1.27	1.21
Modèle de Base	2.53	2.53	2.53

TABLE 4.6 – MAE des modèles entraînés et évalués sur 3 jeux de données pour "Declaration lead time"

Le meilleur modèle est le modèle de Support Vector Machine qui a été obtenu avec les données générées via le TVAE.

Estimation du Release lead time

À présent, nous allons entraîner des modèles afin d'estimer le Release lead time (le lead time que prend la douane pour traiter le dossier de déclaration et émettre un bon à enlever), comme nous l'avons remarqué précédemment dans la phase "Data Understanding", ce lead time est le plus élevé en moyenne et représente un goulot pour le processus d'importation. Son estimation permettra de bien prévoir le lead time total pour chaque importation.
Avant d'entraîner le modèle, nous avons appliqué un ensemble de transformations qui permettent d'améliorer les performances de ces derniers :

— Nous avons appliqué une transformation logarithmique au poids W pour réduire son asymétrie $log(1 + W)$

— Nous avons remplacé les variables nominales par la moyenne du lead time pour chaque valeur unique, cette technique d'encodage est connue sous le nom du "mean target encoding" et permet de donner une représentation numérique efficace aux variables nominales.
— Pour la variable "Regime" pour le régime douanier, on applique un encodage binaire : on remplace "FD" par 0 et "Regime" par 1, tandis que la variable ordinale "Service level" est encodée de la même manière pour ses deux valeurs "Urgent" et "Standard".

Par la suite, en utilisant PyCaret, nous avons lancé l'entraînement de 5 modèles différents : Régression linéaire, SVM, Decision Tree, Gradient Boosted Tree et Random Forest en plus du modèle basique expliqué précédemment, nous avons retrouvé les résultats suivant illustrés dans le tableau 4.7

Modèle	D_1	D_2	D_3
Linear Regression	3.32	4.26	3.05
Decision Tree	3.71	5.38	4.25
Random Forest	3.90	4.24	3.22
Support Vector Machine	4.33	4.56	3.40
Gradient Boosted Trees	4.43	4.27	3.18
Modèle de Base	4.83	4.83	4.83

TABLE 4.7 – MAE des modèles entraînés et évalués sur 3 jeux de données pour "Release lead time"

Selon le tableau 4.7, le meilleur modèle est un modèle simple de régression linéaire entraîné sur le jeu de données D_3 (données originales fusionnées avec les données générées par le TVAE) avec une erreur absolue moyenne de 3.05 jours, ce qui présente une amélioration de 37% de l'erreur de référence (4.83 jours) avec le modèle basique de référence (baseline). De plus, une régression linéaire a l'avantage d'être facile à déployer et à interpréter comparée aux autres modèles.

Estimation du Transport lead time

Pour finir l'étape de modélisation des lead times, nous allons estimer le "Transport lead time" à travers les algorithmes de régression de machine learning, ce lead time représente le délai nécessaire pour transporter la marchandise depuis un des entrepôts douaniers vers une des bases opérationnelles de Schlumberger à Hassi Messaoud. Nous avons appliqué les mêmes transformations déjà effectuées pour estimer "Release lead time" :
— Transformation logarithmique sur le poids pour réduire son asymétrie.
— Encodage des valeurs des variables nominales par la moyenne de y_{transp} correspondante.
— Encodage binaire pour les variables "Service level" et "Regime".

Les résultats obtenus ainsi que l'erreur de référence ont été calculés à travers Python et PyCaret et sont présentés dans le tableau 4.8.

Modèle	D_1	D_2	D_3
Support Vector Machine	1.11	0.97	0.54
Linear Regression	1.39	1.09	0.68
Decision Tree	1.15	1.43	0.63
Random Forest	1.33	1.19	0.63
Gradient Boosted Trees	1.26	1.16	0.67
Modèle de Base	1.43	1.43	1.43

TABLE 4.8 – MAE des différents modèles entraînés et évalués sur 3 jeux de données pour "Transport lead time"

À partir de ces résultats, nous remarquons que le modèle SVM (Support Vector Machine) entrainé sur les données générées par le TVAE avec les données originales D_3 donne le MAE le plus faible de 0.54 jour seulement, ce qui est une amélioration de 62% du modèle basique de la moyenne dont l'erreur est de 1 jour et demi. On remarque d'ailleurs que le jeu de données D_3 permet d'obtenir une erreur MAE inférieur à 1 jour pour tous les modèles testés.

4.4.2 Estimation des coûts de l'importation

Maintenant qu'on a estimé les différents lead times liés au processus d'importation, nous allons à présent s'intéresser à l'estimation des coûts.

Estimation du coût de stockage

Ce coût est engendré par le stockage des marchandises dans les entrepôts douaniers, il est facile de l'estimer, car on a déjà le prix unitaire par jour dans les différents entrepôts et comme on a déjà estimé le Release lead time, le coût total serait donc :

$$\textbf{Coût de stockage} = \textbf{Coût unitaire}_{entrept} * \textbf{Release lead time}$$

Estimation du coût de transport

Les transitaires sont responsables du transport des marchandises entre les point d'entrée et les bases opérationnelles. Les coûts de transports sont déterminés dans les contrats. Afin de calculer le coût, on doit donc avoir 3 paramètres qui sont : *Le point d'entrée*, *La base d'arrivée* et **le poids de l'expédition** afin de choisir le camion adapté.
Les coûts restants sont :

- Le coût lié au traitement de dossier par le transitaire : il est très facile d'avoir le coût exact, car le coût de traitement de dossier est mentionné dans le contrat du transitaire.
- Le coût de dédouanement : il est impossible de prédire ce coût, car il est très variable et il dépend de la réglementation de douane Algérienne.

Le coût total s'écrit donc :

$$C_{total} = C_{transport}(X_i) + C_{stockage}(Y_j) + C_{transitaire}(Z_k) + C_{ddouanement}$$

Avec :
X_i : le camion choisi pour le transport Y_j : Le port ou aéroport d'entrée. Z_k : le transitaire choisi pour la mission de dédouanement.

4.5 Déploiement de la solution

Dans cette section, nous allons présenter la méthodologie de déploiement des modèles entrainés précédemment pour qu'ils soient utilisables par les employées de Schlumberger. D'abord nous allons déployer les modèles entrainés, ensuite nous allons développer un chat bot qui servira comme interface entre le spécialiste import-export de Schlumberger et notre système.

4.5.1 Déploiement des modèles

Nous allons utiliser PyCaret qui offre une méthode pour convertir un modèle entrainé en fichier *Pickle* d'extension "pkl". Par la suite, ces modèles pourront être chargés dans un autre programme ou système à travers une autre méthode présente dans PyCaret.
Pour chaque lead time, nous allons récupérer le modèle entrainé le plus performant : SVM (*Support Vector Machine*) pour "CFD lead time" et "Declaration lead time", le modèle de régression linéaire pour "Release lead time" et un autre SVM pour le "Transport lead time", chaque modèle est par la suite sauvegardé autant que fichier *Pickle* séparément. Il est à noter que le "Transport lead time" peut être estimé plus précisément à la minute près en utilisant les services de Google Maps au lieu d'un modèle.
Nous pouvons par la suite déployer ces fichiers *Pickle* sur le cloud comme AWS (*Amazon Web Services*) ou Microsoft Azure, cependant, vu la confidentialité de ces modèles ainsi que des données, il faudra entamer des discussions avec le département de système d'informations de Schlumberger et les responsables IT de l'entreprise. Pour la suite de ce travail, ces modèles sont déployé localement seulement sur une de nos machines.

4.5.2 Mise en place du système d'estimation

Maintenant que nous avons exporté les modèles d'estimation vers un format de fichier utilisable, nous pouvons mettre en place le système complet qui prend en input une donnée, effectue les différentes transformations, calcule les lead times estimés et propose les scénarios avec les délais et coûts les plus faibles.
Tout d'abord, l'utilisateur fait entrer un ensemble d'information sur l'expédition dont il souhaite estimer les lead times, par exemple le poids, la Business Line concernée, l'entité légale, la quantité, et d'autres informations dont on ne peut pas inférer. Par la suite, l'utilisateur spécifie s'il souhaite comparer les lead times estimés par transitaire (CCA), moyen de transport (MOT) ou bien le port d'entrée (Port of entry), il peut aussi choisir plusieurs dimensions de comparaison à la fois. Selon les dimensions de comparaison choisie, un programme Python génère k exemplaires de l'expédition où k est le nombre de valeurs possibles pour ces dimensions de comparaison, chaque exemplaire comporte une valeur différente par rapport à cette dimension. Pour mieux illustrer cette logique, prenons un cas d'utilisation concret : un spécialiste d'import-export de Schlumberger souhaite estimer et comparer les lead times pour une certaine expédition selon le moyen de transport seulement (la dimension de comparaison est donc la colonne "MOT" uniquement). Il fait donc rentrer quelques informations sur cette expédition comme son poids, la Business Line, l'entité légale, la destination et la quantité. Par la suite, un programme en Python, crée 3 exemplaires de cette ligne de donnée, le premier avec "MOT" correspondant à "Sea freight", un autre à "Air freight" et le troisième à

"Land freight", ces 3 exemplaires sont ensuite passés aux 4 modèles développés précédemment et le résultat pour chaque situation est affiché à l'utilisateur à travers une interface comme un chatbot. Le diagramme de la figure 4.19 présente le fonctionnement de ce système.

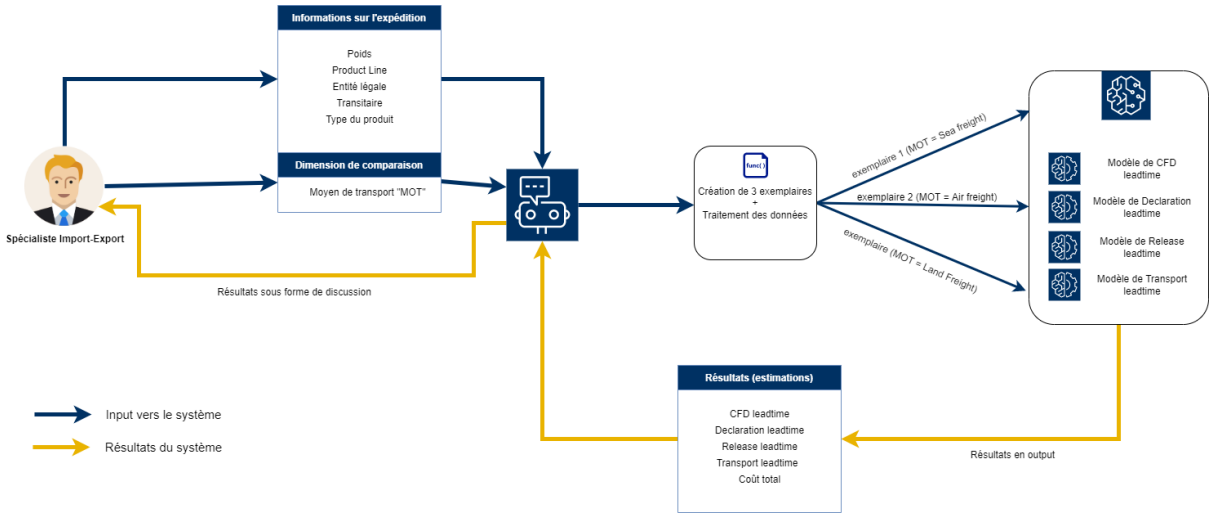


FIGURE 4.19 – Diagramme explicatif du système d'estimation des lead times et des coûts

Dans la figure 4.19, le "MOT" est sélectionné comme dimension de comparaison, cependant, notre système prend en considération plusieurs dimensions de comparaisons en même temps : MOT (*Air freight,Land freight, Sea freight*), CCA (*FD, VAT*) et Port of Entry (*Alger-Port, Skikda-Port, HMD-Airport, Aéroport HMD, etc*)

4.5.3 Implémentation de l'interface du Chat-bot

Dans ce qui suit, nous allons mettre en place une interface interactive sous la forme d'un Chat-bot, qui peut prendre en entrer des informations sur l'expédition ainsi que les dimension de comparaison souhaitée et envoie ces informations au système d'estimation comme illustré dans la figure 4.19. À la fin, le Chat-bot affiche les lead times estimés et le coût total pour chaque scénario possible selon les dimensions de comparaison sélectionnées.
Ce Chat-bot a été développée en utilisant Python et la bibliothèque Streamlit (voir annexe

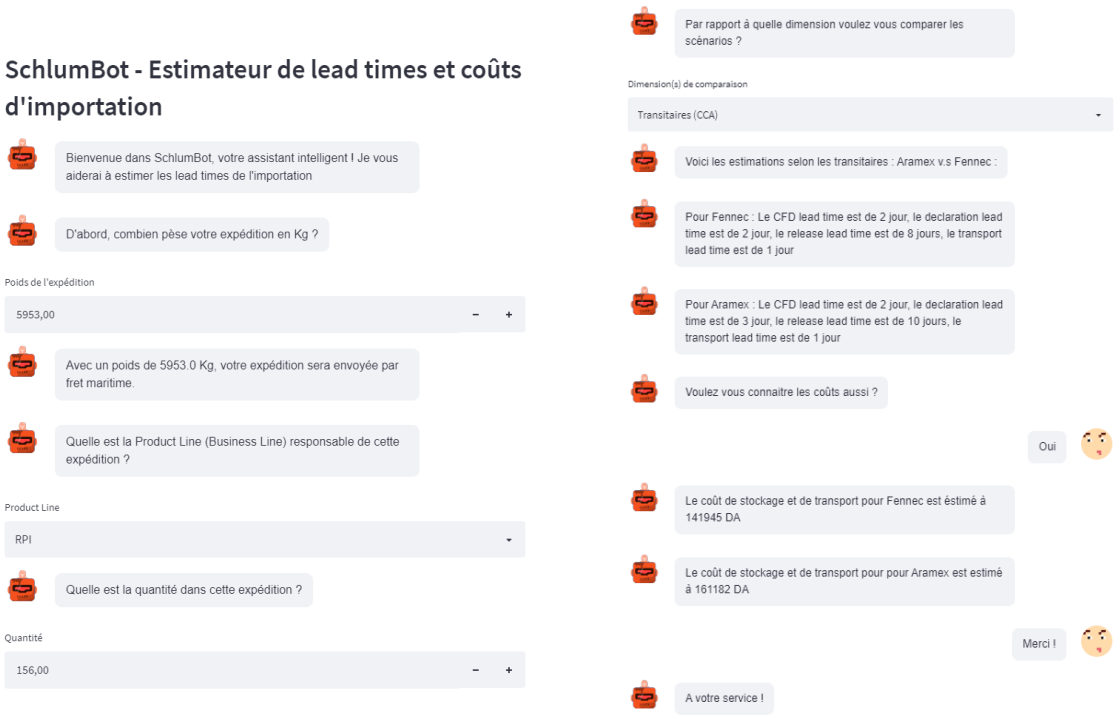


FIGURE 4.20 – Exemple du Chat-bot développé comme interface

Conclusion Générale

Conclusion Générale

À l'heure actuelle de la mondialisation, les lead times liés à l'importation et exportation sont devenues de plus en plus importants et leurs optimisations permettront de créer un avantage concurrentiel.
Lors de notre projet au sein de Schlumberger nous avons appliqué les différentes étapes de la méthodologie CRISP-DM afin de mettre en place un moyen pour estimer ces lead times, la méthodologie s'est déroulée selon 5 étapes.
Nous avons tout d'abord commencé par la compréhension des métiers afin de connaitre en détail le processus d'importation, les différentes parties prenantes et les coûts liés à l'importation. Lors de cette phase, nous avons développé un tableau de bord qui permet d'avoir un suivi des performances des différentes parties prenantes ainsi que le suivi des différents coûts. Nous avons aussi implémenté un système de pénalités pour les transitaires. Ensuite, nous avons déterminé les différentes sources de données ainsi que la définition de chaque colonne afin de les utiliser par la suite dans les différents modèles de Machine Learning.
Après avoir choisi nos sources de données, nous avons préparé l'ensemble des données au format requis par les modèles de Machine Learning en enlevant les données de mauvaise qualité. Par la suite, on est passé à l'étape de modélisation où nous avons testé les différents modèles de Machine Learning, mais, on a constaté que le nombre de données collecté reste insuffisant pour des modèles de Machine Learning, nous avons donc opté pour deux méthodes différentes de synthétisation des données qui sont la simulation et le TVAE : Tabular Variational AutoEncoder. Ces techniques nous ont permis d'améliorer nos modèles et la précision des estimations des différents lead times.
Après avoir estimé l'ensemble des lead times, nous avons déployé notre solution sous forme de Chat-Bot dans le but de faciliter l'utilisation des modèles d'estimation par les différents I/E spécialistes au sein de Schlumberger.

Cette solution permet aux spécialistes d'I/E d'avoir une idée sur les coûts et les lead times liés au processus d'importation. La solution peut être améliorée dans le futur suivant plusieurs axes :

— Amélioration de la précision de l'estimation du temps de transport en utilisant les services de Google Maps qui peuvent nous donner des résultats plus précis.

— Récolter plus de données pour améliorer la fiabilité des modèles au lieu d'employer de la synthétisation de données.

— Intégration les lead times et les coûts de la partie internationale afin de proposer un lead time global. Il est même possible de demander des devis aux entreprises de transport afin d'estimer le coût et le lead time nécessaire pour transporter l'expédition au niveau international.

— L'estimation du coût de dédouanement en analysant en détail les expédition

et les équipements importés dans le but de les classer et estimer les droits et taxes, cette estimation nous permettra d'avoir le coût total lié à l'importation.

Il est à noter aussi que lors du dédouanement, Schlumberger récupère les informations liées aux différentes dates importantes de l'expédition ainsi que les coûts engendrés manuellement en échangeant par email avec le transitaire, puis le spécialiste I/E rempli manuellement la Clearance Portal. Ce processus peut facilement être automatisé à l'aide de la Robotic Process Automation - RPA.
Pour finir, les résultats obtenus dans ce projet aideront le département d'import-export de Schlumberger NAF à mieux planifier leurs expéditions en termes de lead times et de coûts, ce projet a également démontré l'efficacité de l'apprentissage machine dans un contexte industriel où l'environnement est incertain et stochastique par nature, les techniques de génération de données ont aussi prouvé leur utilité dans le cas ou la récolte de données supplémentaires est difficile ou impossible.

Bibliographie

Références

AnyLogic. (2022). *Software for easier supply chain design, analysis, and optimization.* Consulté sur https://www.anylogic.com/blog/software-for-easier-supply-chain-design-analysis-and-optimization/

Azvine, B., Cui, Z., & Nauck, D. D. (2005, juillet). Towards real-time business intelligence. *BT Technology Journal*, *23*(3), 214–225. Consulté sur https://doi.org/10.1007/s10550-005-0043-0 doi: 10.1007/s10550-005-0043-0

Baig, M. M. U., Ali, Y., & Rehman, O. U. (2022, mars). Enhancing Resilience of Oil Supply Chains in Context of Developing Countries. *Operational Research in Engineering Sciences : Theory and Applications*, *5*(1), 69–89. Consulté sur https://oresta.rabek.org/index.php/oresta/article/view/190 (Number : 1) doi: 10.31181/oresta210322091b

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. Consulté sur http://link.springer.com/10.1023/A:1010933404324 doi: 10.1023/A:1010933404324

Castellanos, S. (2021, juillet). Fake It to Make It : Companies Beef Up AI Models With Synthetic Data. *Wall Street Journal*. Consulté sur https://www.wsj.com/articles/fake-it-to-make-it-companies-beef-up-ai-models-with-synthetic-data-11627032601

Chaillou, T., & Dugué, N. (2021). Introduction [Analyser les données de l'entreprise avec Power BI]. Consulté sur https://geainfolemans.github.io/PowerBI/co/module__2.html

Cheng, M., Fang, F., Kinouchi, T., Navon, I. M., & Pain, C. C. (2020). Long lead-time daily and monthly streamflow forecasting using machine learning methods. *Journal of Hydrology*, *590*, 125376. (Publisher : Elsevier)

Cohen, M. A., & Huchzermeier, A. (1999). Global supply chain management : A survey of research and applications. In F. S. Hillier, S. Tayur, R. Ganeshan, & M. Magazine (Eds.), *Quantitative Models for Supply Chain Management* (Vol. 17, pp. 669–702). Boston, MA : Springer US. Consulté sur http://link.springer.com/10.1007/978-1-4615-4949-9_21 (Series Title : International Series in Operations Research & Management Science) doi: 10.1007/978-1-4615-4949-9_21

Colicchia, C., Dallari, F., & Melacini, M. (2010, octobre). Increasing supply chain resilience in a global sourcing context. *Production Planning & Control*, *21*(7), 680–694. Consulté sur https://doi.org/10.1080/09537280903551969 (Publisher : Taylor & Francis _eprint : https ://doi.org/10.1080/09537280903551969) doi: 10.1080/09537280903551969

Direction Générale des Douanes. (2022). *Procédure de dédouanement.* Consulté sur https://www.douane.gov.dz/spip.php?rubrique30

Foley, , & Guillemette, M. G. (2010, octobre). What is Business Intelligence ? *International Journal of Business Intelligence Research*

(IJBIR), *1*(4), 1–28. Consulté sur https://www.igi-global.com/article/business-intelligence/www.igi-global.com/article/business-intelligence/47193 (Publisher : IGI Global) doi: 10.4018/jbir.2010100101

Gartner. (2021, juillet). *By 2024, 60% of the data used for the development of AI and analytics projects will be synthetically generated.* Consulté sur https://blogs.gartner.com/andrew_white/2021/07/24/by-2024-60-of-the-data-used-for-the-development-of-ai-and-analytics-projects-will-be-synthetically-generated/

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014, juin). *Generative Adversarial Networks* (Rapport technique N° arXiv :1406.2661). arXiv. Consulté sur http://arxiv.org/abs/1406.2661 (arXiv :1406.2661 [cs, stat] type : article) doi: 10.48550/arXiv.1406.2661

Gyulai, D., Pfeiffer, A., Nick, G., Gallina, V., Sihn, W., & Monostori, L. (2018). Lead time prediction in a flow-shop environment with analytical and machine learning approaches. *IFAC-PapersOnLine*, *51*(11), 1029–1034. (Publisher : Elsevier)

Hayya, J., Ramasesh, R., Tyworth, J., Kim, J., & Sun, D. (2013). "JIT" delivery with stochastic lead time. *The Journal of the Operational Research Society*, *64*(1), 97–105. Consulté sur http://www.jstor.org/stable/23355375 (Publisher : Palgrave Macmillan Journals)

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278–282). IEEE.

Inmon, W. H. (2002). *Building the data warehouse* (3rd ed éd.). New York : J. Wiley.

Intelligence, M. (2022). *Oilfield Services Market | 2022 - 27 | Industry Share, Size, Growth - Mordor Intelligence.* Consulté sur https://www.mordorintelligence.com/industry-reports/global-oil-field-services-market-outlook-industry

Ioannou, G., & Dimitriou, S. (2012, octobre). Lead time estimation in MRP/ERP for make-to-order manufacturing systems. *International Journal of Production Economics*, *139*(2), 551–563. Consulté sur https://www.sciencedirect.com/science/article/pii/S0925527312002216 doi: 10.1016/j.ijpe.2012.05.029

James, G., Witten, D., Hastie, T., & Tibshirani, R. (Eds.). (2013). *An introduction to statistical learning : with applications in R* (N° 103). New York : Springer. (OCLC : ocn828488009)

Jeremy. (2021, janvier). *Power BI, la solution de Business Intelligence de Microsoft.* Consulté sur https://datascientest.com/power-bi

KDnuggets. (2021). *Data Science vs Business Intelligence, Explained.* Consulté sur https://www.kdnuggets.com/data-science-vs-business-intelligence-explained.html/ (Section : 2021 Feb Tutorials, Overviews)

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit : the definitive guide to dimensional modeling* (Third edition éd.). Indianapolis, IN : John Wiley & Sons, Inc. (OCLC : ocn840431951)

KPMG. (2021). *Six key trends impacting global supply chains in 2022 - kpmg global.* Consulté sur https://home.kpmg/xx/en/home/insights/2021/12/

six-key-trends-impacting-global-supply-chains-in-2022.html
Lingitz, L., Gallina, V., Ansari, F., Gyulai, D., Pfeiffer, A., Sihn, W., & Monostori, L. (2018, janvier). Lead time prediction using machine learning algorithms : A case study by a semiconductor manufacturer. *Procedia CIRP*, *72*, 1051–1056. Consulté sur https://www.sciencedirect.com/science/article/pii/S2212827118303056 doi: 10.1016/j.procir.2018.03.148
Lowe, D. (2002). *The dictionary of transport and logistics.* London : Kogan Page. (OCLC : ocm49204234)
Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., ... Flach, P. (2021, août). CRISP-DM Twenty Years Later : From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3048–3061. (Conference Name : IEEE Transactions on Knowledge and Data Engineering) doi: 10.1109/TKDE.2019.2962680
McKinsey. (2022). *Overcoming global supply chain challenges.* Consulté sur https://www.mckinsey.com/industries/travel-logistics-and-infrastructure/our-insights/overcoming-global-supply-chain-challenges#:~:text=the%20three%20critical%20challenges%20facing,a%20climate%20of%20persistent%20unpredictability
Mourtzis, D., Doukas, M., Fragou, K., Efthymiou, K., & Matzorou, V. (2014, janvier). Knowledge-based Estimation of Manufacturing Lead Time for Complex Engineered-to-order Products. *Procedia CIRP*, *17*, 499–504. Consulté sur https://www.sciencedirect.com/science/article/pii/S2212827114003394 doi: 10.1016/j.procir.2014.01.087
Nguyen, T., Schiefer, J., & Tjoa, A. M. (2005, janvier). Sense & response service architecture (SARESA) : an approach towards a real-time business intelligence solution and its use for a fraud detection application. (Pages : 86) doi: 10.1145/1097002.1097015
Nima, P. (2019). *Quora Insincere Questions Classification.*
Oil 2021 (Rapport technique). (2021). IEA, Paris : IEA.
Panoply. (s. d.). *Data Mart vs. Data Warehouse.* Consulté sur https://panoply.io/data-warehouse-guide/data-mart-vs-data-warehouse/
Parmenter, D. (2020). *Key performance indicators : developing, implementing, and using winning KPIs* (Fourth edition éd.). Hoboken, New Jersey : John Wiley & Sons, Inc.
Pine, D. (2021, septembre). *Types of Dashboards : Strategic, Operational & Analytical.* Consulté sur https://www.datapine.com/blog/strategic-operational-analytical-tactical-dashboards/
Ponomarov, S. Y., & Holcomb, M. C. (2009, janvier). Understanding the concept of supply chain resilience. *The International Journal of Logistics Management*, *20*(1), 124–143. Consulté sur https://doi.org/10.1108/09574090910954873 (Publisher : Emerald Group Publishing Limited) doi: 10.1108/09574090910954873
Sarder, M. (2021). *Logistics Transportation Systems.* USA : Elsevier.
Schlumberger Ltd. (2021). *2021 Annual Report* (Rapport technique). Schlumberger Ltd.
Shipper, D. (2021). *FCL vs LCL : Comment choisir la meilleure expédition ?* Consulté sur https://sourcing.docshipper.com/sourcing/fcl-vs-lcl-meilleure-expedition/
Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., & Ganguli, S. (2015,

novembre). *Deep Unsupervised Learning using Nonequilibrium Thermodynamics* (Rapport technique N° arXiv :1503.03585). arXiv. Consulté sur http://arxiv.org/abs/1503.03585 (arXiv :1503.03585 [cond-mat, q-bio, stat] type : article) doi: 10.48550/arXiv.1503.03585

Sterman, J. D. (2009). *Business dynamics : systems thinking and modeling for a complex world* (Nachdr. éd.). Boston : Irwin/McGraw-Hill.

TIOBE Index. (2022). Consulté sur https://www.tiobe.com/tiobe-index/

Tsolaki, K., Vafeiadis, T., Nizamis, A., Ioannidis, D., & Tzovaras, D. (2022, février). Utilizing machine learning on freight transportation and logistics applications : A review. *ICT Express*. Consulté sur https://www.sciencedirect.com/science/article/pii/S2405959522000200 doi: 10.1016/j.icte.2022.02.001

Unzueta, D. (s. d.). *How to Generate Tabular Data Using CTGANs | by Diego Unzueta | Towards Data Science.* Consulté sur https://towardsdatascience.com/how-to-generate-tabular-data-using-ctgans-9386e45836a6

Vanajakumari, M., Sun, H., Jones, A., & Sriskandarajah, C. (2022). Supply chain planning : A case for hybrid cross-docks. , *108*. Consulté sur https://www.sciencedirect.com/science/article/pii/S0305048321001948 doi: 10.1016/j.omega.2021.102585

Weflen, E., MacKenzie, C. A., & Rivero, I. V. (2022, janvier). An influence diagram approach to automating lead time estimation in Agile Kanban project management. *Expert Systems with Applications*, *187*, 115866. Consulté sur https://www.sciencedirect.com/science/article/pii/S0957417421012252 doi: 10.1016/j.eswa.2021.115866

Welsing, M., Maetschke, J., Thomas, K., Gützlaff, A., Schuh, G., & Meusert, S. (2020). Combining Process Mining and Machine Learning for Lead Time Prediction in High Variance Processes. In *Congress of the German Academic Association for Production Technology* (pp. 528–537). Springer.

Weng, L. (2018). From autoencoder to beta-vae. *lilianweng.github.io*. Consulté sur https://lilianweng.github.io/posts/2018-08-12-vae/

Weng, L. (2021, juillet). *What are Diffusion Models ?* Consulté sur https://lilianweng.github.io/posts/2021-07-11-diffusion-models/ (Section : posts)

World Bank. (2020). *National accounts data - World Bank / OECD.* http://data.worldbank.org/data-catalog/world-development-indicators.

World Bank Group. (2020, novembre). *Algeria Economic Monitor, Fall 2020 : Navigating the COVID-19 Pandemic, Engaging Structural Reforms* (Rapport technique). Washington, DC : World Bank. Consulté sur https://openknowledge.worldbank.org/handle/10986/35058 (Accepted : 2021-01-27T16 :23 :05Z)

World Customs Organization. (2018). *Glossary of international customs terms.* WCOOMD. Consulté sur http://www.wcoomd.org

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019a). *Modeling tabular data using conditional gan.* arXiv. Consulté sur https://arxiv.org/abs/1907.00503 doi: 10.48550/ARXIV.1907.00503

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019b, octobre). *Modeling Tabular data using Conditional GAN* (Rapport technique N° arXiv :1907.00503). arXiv. Consulté sur http://arxiv.org/abs/1907.00503 (arXiv :1907.00503 [cs, stat] type : article) doi: 10.48550/arXiv.1907.00503

Annexes

Annexe A

Tests statistiques

Cette annexe a pour but de présenter les différents tests statistiques employés dans ce projet, nous présenterons le test de Shapiro-Wilk pour tester la normalité, le test de Levene pour tester l'homoscédasticité entre plusieurs groupes, ensuite nous présenterons l'ANOVA classique et sa variante non paramétrique : l'ANOVA de Kruskal-Wallis.

A.1 Le test de normalité de Shapiro-Wilk

Afin de tester si une série de données est distribuée normalement, on dispose de plusieurs tests statistiques, les plus connus sont le test de Kolmogorov-Smirnov (1939) et le test de Shapiro-Wilk (1965). Le test de Kolmogorov-Smirnov (KS) est un test qui permet de tester si un échantillon suit une certaine loi de distribution déterminée, tandis que le test de Shapiro-Wilk (SW) est employé pour tester si un échantillon suit une loi normale, ce dernier permet de ne pas spécifier la moyenne et l'écart type de loi normale qu'on veut tester au contraire du test de Kolmogorov-Smirnov.
Le test de Shapiro-Wilk test l'hypothèse suivante :
- $\mathbb{H}_0$: Les données sont distribuées normalement.
- $\mathbb{H}_1$: Les données ne proviennent pas d'une distribution normale.

Pour cela, on calcule la statistique suivante :

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{(x_i - \bar{x})^2}$$

Les x_i sont les données ordonnées et les coefficients a_i sont calculés par l'opération matricielle suivante :

$$(a_1, a_2, ..., a_n) = \frac{m^t V^{-1}}{(m^t V^{-1} V^{-1} m)^{1/2}}$$

Avec m étant le vecteur des espérances, V est la matrice des covariances. En pratique, on utilise un outil ou une bibliothèque comme SciPy qui permet d'effectuer ce test aisément et nous donne les p-values (valeurs p), si la valeur p est inférieure au seuil de risque statistique α (qu'on prend 0.05 généralement) on rejette l'hypothèse nulle et les données ne sont pas distribuées normalement sinon ils le sont.

A.2 Test d'homoscédasticité de Levene

Le test de Levene permet de tester si deux ou plusieurs groupes différents de données ont la même variance (l'homoscédasticité). Ce test est utilisé principalement pour vérifier une des hypothèses de l'ANOVA (présenté prochainement) qui concerne l'homoscédasticité ou l'égalité des variances.
L'hypothèse testée est la suivante :
- $\mathbb{H}_0 : \sigma_1 = \sigma_2 = \ldots = \sigma_k$
- $\mathbb{H}_1 : \sigma_i = \sigma_j$ pour au moins une paire i, j.

Mathématiquement, on calcule la statistique suivante :

$$W = \frac{(N-k)}{(k-1)} \cdot \frac{\sum_{i=1}^{k} N_i (Z_{i\cdot} - Z_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Z_{i\cdot})^2}$$

N est le nombre total de données, k est le nombre de groupes, Z_{ij} est la valeur de la variable en question de la donnée j dans le groupe i centré par la moyenne de cette variable. La statistique W suit approximativement une loi de Fisher $\mathcal{F}(1 - \alpha, N - k, k - 1)$, il suffit donc de vérifier la valeur de W et la valeur de la table correspondante au seuil statistique, ou bien comparer la valeur p donnée par Scipy ou autre pour accepter ou rejeter l'hypothèse nulle.

A.3 ANOVA Classique

L'analyse des variances *(Analysis of Variances - ANOVA)* est une méthode statistique qui permet de déterminer si les moyennes de plusieurs groupes proviennent de la même population. Cette méthode est utilisée dans plusieurs contextes, elle est employée par exemple pour déterminer l'effet d'une variable catégorique (nominal ou ordinal) sur une variable numérique continue. Le test d'ANOVA classique à un facteur teste l'hypothèse suivante :
- $\mathbb{H}_0$: Les moyenne ne sont pas significativement différentes entre les groupes $\mu_1 = \mu_2 = \ldots = \mu_k$.
- $\mathbb{H}_1$: Au moins une moyenne est différente.

Cependant, avant d'appliquer le test d'ANOVA, il faut vérifier un ensemble de conditions relatives aux données : Les données (mesures) doivent être indépendantes entre elles. Les résidus doivent être distribués selon une distribution normale. Les variances entre les différents groupes des résidus analysés doivent être égale (homoscédasticité des groupes).
Le principe est de décomposer les valeurs observées (chaque valeur) en la somme de la moyenne arithmétique de la donnée, la différence entre les moyennes de groupes et la moyenne totale, et l'écart entre la valeur et la moyenne. Cela nous donne pour une valeur :

$$x_{ij} = \bar{x} + (\bar{x}_i - \bar{x}) + (x_{ij} - \bar{x}_i)$$

La différence entre la moyenne du groupe ($\bar{x}_i$) et la moyenne totale ($\bar{x}$) représente l'effet α_i de la variable catégorique sur la variable numérique (pour un certain groupe i). La variance est ensuite décomposée en calculant la somme des carrées entre groupe $\mathrm{SSD_B}$ et la somme des résidus $\mathrm{SSD_W}$:

$$\mathrm{SSD_B} = \sum_{i=1}^{k} \sum_{j=1}^{n} (\bar{x}_i - \bar{x})^2 = n \sum_{i=1}^{k} (\bar{x}_i - \bar{x})^2$$

$$\mathrm{SSD_W} = \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2$$

On divise par la suite ces sommes par leurs degrés de liberté et on prend leur ratio, on obtient par cela :

$$F = \frac{\frac{\mathrm{SSD_B}}{(k-1)}}{\frac{\mathrm{SSD_W}}{(N-k)}} \sim F_{\alpha=0.05}(k-1, N-k)$$

La statistique F suit une distribution de Fisher aux degrés de liberté $(N-k, k-1)$ qu'on peut comparer à la valeur critique dans la table de Fisher et par la suite accepter ou rejeter l'hypothèse nulle, on peut aussi utiliser les p-valeurs. L'effet d'une variable catégorique sur une variable numérique continue peut être quantifié par la suite en calculant le coefficient éta carré (η^2) donnée par :

$$\eta^2 = \frac{\mathrm{SSD_B}}{\mathrm{SSD_W} + \mathrm{SSD_B}}$$

Quand éta carré est supérieur à 0.2, cela veut dire que la variable a un effet considérable sur la variable numérique, tandis que si elle est proche de 0, la variable catégorique n'a aucune influence sur les valeurs.
Quand la condition de normalité et d'homogénéité des variances n'est pas vérifiée, l'ANOVA n'est plus fiable et perd sa puissance statistique, dans ce cas, on emploie une variante non paramétrique de l'ANOVA qui est l'ANOVA de Kruskal-Wallis.

A.4 ANOVA de Kruskal-Wallis

Le test ANOVA de Kruskal Wallis est une variante non paramétrique de l'ANOVA classique, elle a été développée par les statisticiens William Kruskal et Wilson Allen Wallis. Ce test permet d'appliquer une ANOVA quand la condition de distribution normale des données et des résidus n'est pas vérifiée. Elle peut aussi être utilisée quand d'autres conditions ne sont pas vérifiées, par exemple si la variance n'est pas homogène entre les groupes comparés (l'hétéroscédasticité des données). Ce texte se base sur les rangs (ordre des données) au lieu des valeurs mesurées. En se basant sur les rangs, cela permet de ne pas prendre en considération la distribution des données, ce qui donne un avantage important comparé à l'ANOVA classique (paramétrique).
Ce test permet de tester l'hypothèse suivante qui est similaire aux hypothèses de l'ANOVA mais avec des médianes au lieu de moyennes :

- $\mathbb{H}_0$: Les médianes ne sont pas significativement différentes entre les groupes $m_1 = m_2 = \ldots = m_k$.
- $\mathbb{H}_1$: Au moins une médiane est différente.

Le test de Kruskal Wallis peut être effectué en suivant ces étapes :

- Classer les données par ordre en combinant tous les groupes.
- Placer le rang de chaque point de données.
- Additionner les rangs des groupes dans le même point
- Calculer la statistique H donnée par :

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{C} \frac{R_i^2}{n_i} - 3(N+1)$$

N étant le nombre de données, C est le nombre de groupes, R_i est la somme des rangs dans le $i$ème groupe, n_i est la taille du $i$ème groupe.
On compare cette valeur à une distribution du 2 (chi-2) à $C-1$ degrés de liberté, on peut aussi utiliser les valeurs p et les comparer au seuil statistique α : Si $p > \alpha$ on accepte l'hypothèse nulle et donc les médianes ne sont pas significativement différentes entre les groupes, dans le cas contraire, on rejette l'hypothèse nulle et les médianes sont significativement différente dans ce cas. En pratique, la bibliothèque Pingouin de Python offre une méthode pour lancer ce test en une seule ligne de code et avoir les p-valeurs et d'autres mesures. À partir de cela, il est possible de déterminer l'effet d'une variable catégorique sur une variable numérique continue comme l'ANOVA classique, cependant il n'existe pas une formule unique dans le cas d'une ANOVA de Kruskal-Wallis pour déterminer cet effet η^2, certains chercheurs ont proposé des formules à partir de la statistique H calculée tel que la formule suivante :

$$\eta^2(H) = \frac{H - C + 1}{N - C}$$

Annexe B

Détails sur le Machine Learning

Dans cette annexe, nous allons compléter la présentation des algorithmes de Random Forest, Gradient Boosted Trees et nous allons présenter le problème du sur-apprentissage (overfitting) et de validation croisée (cross-validation), et pour finir nous allons présenter la divergence de Kullback-Leibler utilisé pour comparer deux distributions et qu'on utilise pour évaluer la qualité des données générée dans le quatrième chapitre.

B.1 Algorithmes du Random Forest et du Gradient Boosted Trees

Dans ce qui suit, nous présenterons le déroulement des algorithmes du Random Forest et du Gradient Boosted Trees en détails.

B.1.1 Algorithme du Random Forest

Nous avons déjà présenté cet algorithme brièvement dans le chapitre 2, pour plus de détails, l'algorithme se déroule de la manière suivante (Breiman, 2001) :

1. On génère B échantillons de données à travers le bootstraping.
2. On entraîne B arbres de décisions avec un nombre aléatoire des variables exogènes disponibles (en général si on a p variables exogènes (dimension/facteurs), chaque arbre sera entraîné sur $\sqrt{p}$ de variables).
3. On agrège le résultat par la moyenne de ces arbres :

$$\hat{f}_{\text{moy}} = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}$$

Cette procédure permet d'entraîner un modèle fiable et robuste avec une variance réduite.

B.1.2 Algorithme du Gradient Boosted Trees

Cet algorithme fut aussi présenté dans le deuxième chapitre brièvement, il fonctionne de la manière suivante (James et al., 2013) :

1. Pour la première itération, on pose $r_i = y_i$ pour toutes les données i.
2. On choisit un nombre B d'arbres à entraîner :

3. Pour $b = 1, 2, 3, \ldots, B$:
   (a) On entraîne un arbre $\hat{f}^b$ avec d coupes ($d + 1$ feuilles).
   (b) On met à jour le modèle en ajoutant cet arbre multiplié par un coefficient $\lambda < 1$:

$$\hat{f}(x) = \hat{f}(x) + \lambda \hat{f}^b(x)$$

   (c) Mettre à jour les résidus :

$$r_i = r_i - \lambda \hat{f}^b(x_i)$$

4. calculer le modèle final :

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x)$$

B.2 Sur-apprentissage et validation croisée

Le sur-apprentissage, connu aussi sous le nom d'overfitting, est un phénomène très fréquent en Machine Learning, un modèle est en overfitting s'il se focalise énormément sur les données fournies et n'arrive pas à généraliser son apprentissage (il n'arrive pas à découvrir la tendance générale des données). Formellement, l'erreur sur un jeu de donnée d'évaluation (test set) devient plus grande après un certain nombre d'itérations que l'erreur calculée sur le jeu de donnée d'entraînement (train set) (B.1).

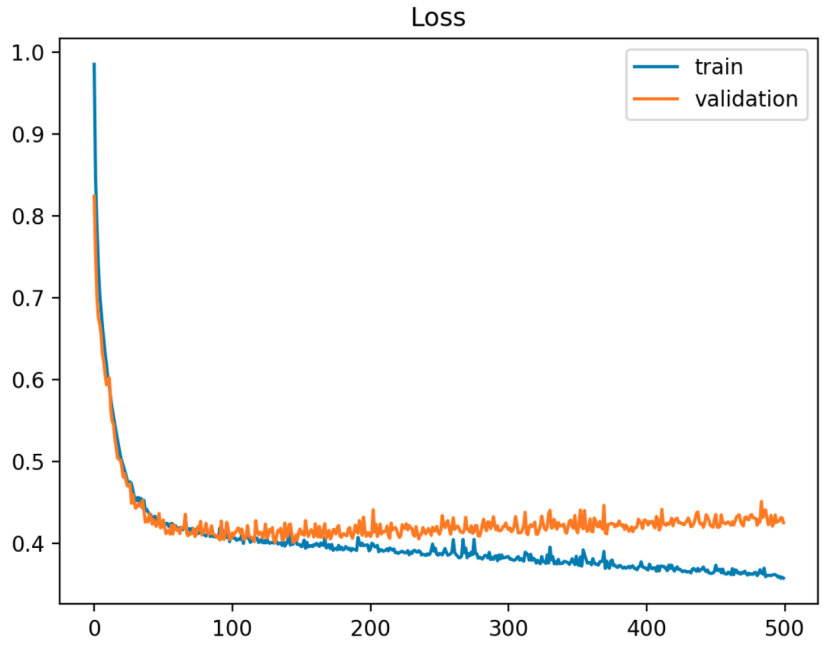


FIGURE B.1 – Erreur d'entraînement et de validation (évaluation) dans une situation d'overfitting

En général, un modèle comporte un biais qui représente la tendance de ce modèle en vers les données et une variance qui présente comment le modèle réagi à une

fluctuation dans les données d'entrée, l'erreur de validation est aussi décomposée en trois termes : le biais $b(x)$, la variance $v(x)$ et une erreur irréductible σ_x :

$$\mathrm{Err}_{\mathrm{test}} = b(x) + v(x) + \sigma_x$$

Réduire le biais augmente généralement la variance et vice-versa, ce phénomène est connu sous le nom de *Bias-Variance trade-off*. Quand le biais est faible, mais que la variance est élevée, on tombe dans une situation de sur-apprentissage, le modèle est donc instable et une simple variation dans les données peut altérer sa performance considérablement. Un bon modèle est un modèle qui minimise le biais et la variance.
La validation croisée (*Cross-Validation*) est une méthode de validation qui permet de détecter des situations d'overfitting et connaître si un modèle est stable ou pas, l'idée est de diviser un jeu de données en k partitions (*k-Folds*) égales et entrainer k fois le même modèle sur $k-1$ des partitions (folds), pour chaque itération, on évalue le modèle sur la partition qui n'a pas été utilisé pour l'entrainement. On calcule l'erreur du modèle pour chaque itération et à la fin, on prend la moyenne des erreurs sur toutes les itérations, ceci nous donne une mesure plus fiable et plus robuste de l'erreur du modèle. Si l'erreur change significativement entre chaque itération, l'écart type de l'erreur est donc élevée, ce qui signifie que le modèle est instable et qu'on est présence d'une situation d'overfitting, la figure B.2 illustre cette méthodologie.

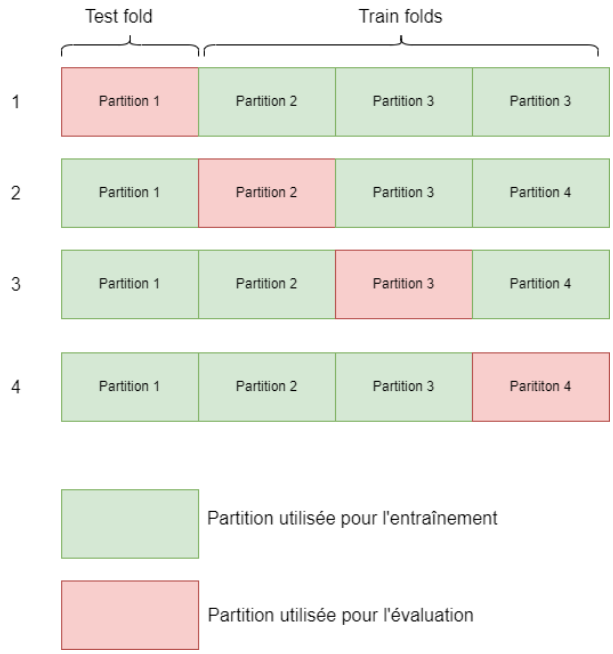


FIGURE B.2 – Diagramme explicatif de la méthodologie de validation croisée (cross-validation)

Pour mieux illustrer cette situation, nous avons entraîné un simple modèle de régression linéaire sur les 3 jeux de données différents (avec les données originales D_1, avec les données originales fusionnées avec des données simulées D_2, avec des données originales fusionnées avec des données générées par le TVAE), la taille de chaque jeu de données est plus grande que l'autre : taille(D_1)taille(D_2)taille(D_3)

Jeu de données	D_1	D_2	D_3
Taille (de lignes)	186	686	686
Erreur moyenne	12.02	8.09	7.7
Écart-type de l'erreur	1.89	0.23	1.0

TABLE B.1 – Moyenne et écart-type de l'erreur avec des jeux de données de tailles différentes

B.2.1 Divergence de Kullback-Leibler

La divergence de Kullback-Leibler est une mesure statistique de dissimilarité entre deux distributions P et Q, cette mesure est définie dans le cas discret par la formule suivante :

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

Dans le cas continue, la somme est remplacée par une intégrale sur toutes les valeurs possibles. La divergence KL est donc l'espérance de la différence des logarithmes de P et Q, quand P est très différent de Q, D_{KL} est élevé, quand P est similaire à Q alors D_{KL} est faible et quand $P = Q$, D_{KL} s'annule.
Cette mesure est utilisée dans ce travail pour mesurer la similarité des données généré avec les données réelles (on mesure la similarité de leurs distributions) ainsi que dans la définition de la fonction d'erreur des réseaux VAE dans l'annexe C.

Annexe C

Génération des données avec TVAE

Dans cette annexe, nous allons présenter le principe du fonctionnement du réseau de neurones artificiel VAE (*Variational Auto-Encoders*) ainsi que sa variante pour les données tabulaires TVAE qui nous a permis de générer des données similaires à 80% aux données réelles. Nous commencerons d'abord par présenter le modèle de type Auto-Encoders (AE), ensuite, nous présenterons le fonctionnement des "Variational Auto-Encoders" (VAE) mais nous donneront qu'une simple présentation mathématique sans entrer dans les détails vu que ce domaine dérive de l'inférence Bayésienne qui ne sera pas présentée dans ce travail.

C.1 Auto-Encoders

Les réseaux de neurones de type "Auto-Encoders" sont des modèles classiques introduit dans les années 1980, leur but est d'apprendre une présentation des données en entrée x dans un espace latent avec une dimension inférieur à la dimension des données. Ils peuvent donc être utilisés comme un moyen pour réduire la dimensionnalité des données comme l'analyse à composante principale (ACP). Sauf que ces réseaux permettent de prendre en compte des relations complexes et non linéaires et leur fonctionnement est différent du principe de l'ACP.
Un "Auto-Encoder" est un réseau composé de deux blocs : un réseau encodeur $E(x)$ et un autre réseau décodeur, le but de l'encodeur est d'apprendre la meilleure représentation de x, dans l'espace latent et donc produire un vecteur z (appelé vecteur latent) dans un espace de dimension inférieur aux dimensions des données initiales x et donc on peut écrire :

$$z = E(x)$$

Le deuxième bloc, est un réseau décodeur $D(z)$, qui permet de reconstruire x à partir du vecteur latent z. Cette reconstruction n'est pas parfaite, car durant la première étape, la réduction de dimension fait perdre de l'information, on obtient donc une donnée (vecteur) $\hat{x}$ qui se rapproche du vecteur initial x. Le but est d'améliorer cette reconstruction durant l'apprentissage en minimisant l'erreur de reconstruction $L(x, \hat{x})$ qui est simplement une certaine mesure de dissimilarité entre x et $\hat{x}$. Généralement, l'erreur peut être la norme p de la différence entre le vecteur initiale x et la reconstruction $\hat{x}$

$$L(x, \hat{x}) = ||x - \hat{x}||_p$$

Sachant que $\hat{x} = D(z)$ et $D(z) = E(x)$ alors $\hat{x} = D(E(x))$ et on peut écrire l'erreur

de reconstruction :

$$L(x, \hat{x}) = ||x - D(E(x))||_p$$

La figure

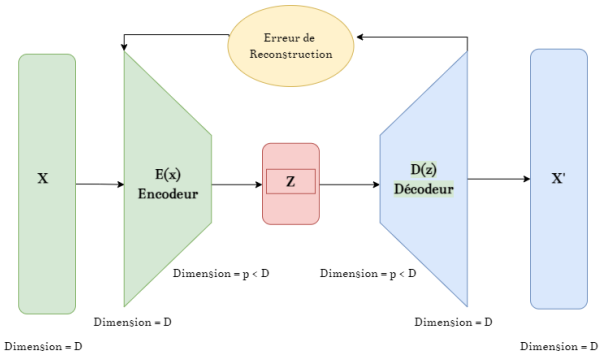


FIGURE C.1 – Diagramme explicatif du fonctionnement d'un réseau de type "Auto-Encoder"

C.2 Variational Auto-Encoders

Bien que les réseaux "Auto-Encoders" ne permettent pas de générer de nouvelles données, mais seulement apprendre une représentation latente z de la donnée x avec une dimension inférieure à la dimension de x, la compréhension du fonctionnement et de l'intuition derrière ces modèles est essentiel pour comprendre comment un réseau VAE (Variational Auto-Encoder) permet de générer des données.
Théoriquement, pour générer des données à partir d'un réseau "Auto-Encoder", il suffit de prendre un échantillon de l'espace latent G_z, en réalité, ceci est impossible, car cette espace n'est pas structuré et n'est pas forcément continue. En effet, il n'existe aucune contrainte qui impose une certaine structure à cette espace et assure sa continuité : si on prend deux données z_1 et Z_2 proche dans cette espace et on produit deux données $\hat{x_1}$ et $\hat{x_2}$, rien n'assure que ces deux données générées seront similaires ou proches, car cette espace n'est pas structuré et pas forcément continue.
L'idée est d'imposer une contrainte pour structurer cet espace latent G_z et assurer sa continuité, ce qui nous permettra de générer de nouvelles données à partir de cet espace avec le même caractéristique des données réelles qu'on dispose. Pour cela, le réseau encodeur E est modifié de façon à produire une distribution (à priori) $p(z)$, de plus, nous allons imposer que cette distribution soit une distribution normale de moyenne nulle est de variance égale à l'identité I, le décodeur, quant à lui, va produire une distribution $p(x|z)$ qui est similaire à une loi normale centrée $\mathbb{N}(0, I)$, cependant, la moyenne dans ce cas sera une fonction du vecteur latent $f(z)$ tandis que l'écart-type sera égal à une constante multipliée par l'identité cI :

$$p(z) = \mathcal{N}(0, I)$$
$$p(x|z) = \mathcal{N}(f(z), cI)$$

Par la suite, on s'intéresse à calculer la distribution postérieure $p(z|x)$ (la probabilité d'obtenir z pour une certaine donnée x), la règle de Bayes nous permet d'obtenir une formulation de cette distribution :

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} = \frac{p(x|z)p(z)}{\int_x p(x|u)p(u)}$$

Le calcul de $p(z|x)$ est en fonction de $p(x|z)$ ainsi que $p(z)$ qu'on connait déjà, cependant, $p(x)$ est pratiquement impossible à calculer analytiquement, car il faudra calculer la distribution sur toutes les données x (d'où l'intégrale sur x). Pour cela, on procède à la méthode d'inférence variationnelle qui permet d'approximer numériquement des distributions complexes (comme $p(x)$) : l'idée est d'utiliser une distribution $q_x(z;\nu)$ dont on ne connaît pas les paramètres ν. Afin de connaitre ces paramètres ν, on utilise la divergence de Kullback-Leibler D_{KL} qu'on a déjà présenté dans l'annexe B.

De ce fait, la fonction d'erreur du VAE contient un nouveau terme en plus de l'erreur de reconstruction qui est la divergence KL entre la distribution variationnelle $q_x(z;\nu)$ et la distribution qu'on cherchait à calculer $p(z|x)$:

$$L(x,\hat{x}) = ||x - D(E(x))||_p + D_{KL}[q_x(z;\nu); p(z|x)]$$

La minimisation de $L(x,\hat{x})$ est mathématiquement complexe et ne sera pas détaillé dans ce travail, mais l'intuition est la suivante : minimiser cette erreur permet de connaître les paramètre ν pour déterminer la distribution $q_x(z;\nu)$, cette distribution est similaire à la distribution $p(z|x)$ qu'on cherchait au début, car la divergence KL est minimale après la minimisation de $L(x,\hat{x})$, or nous n'avons plus besoin d'utiliser la règle de Bayes et de calculer l'intégrale intraitable pour connaître la distribution de z pour chaque x.

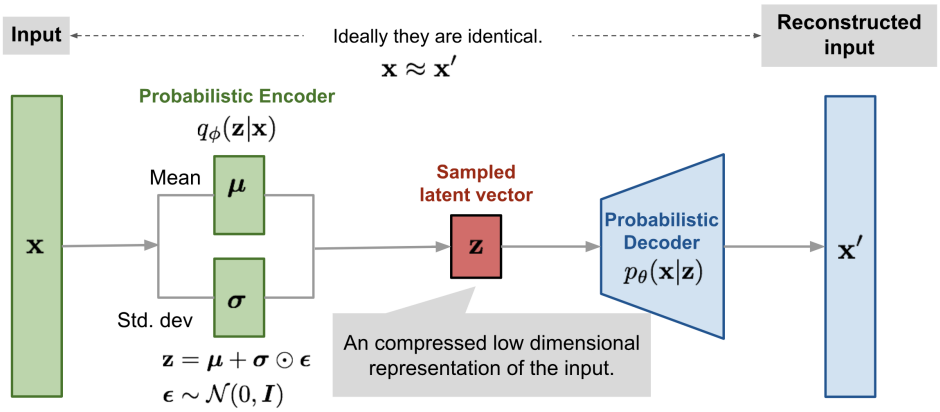


FIGURE C.2 – Diagramme explicatif du fonctionnement d'un réseau de type "Variational Auto-Encoder" (Weng, 2018)

C.3 Tabular Variational Auto-Encoders (TVAE)

Le modèle qu'on utilise pour générer des données tabulaires dans ce travail est une légère variation du VAE traditionnel qui permet de prendre en compte la nature des données tabulaires proposé par (Xu, Skoularidou, Cuesta-Infante, & Veeramachaneni, 2019a), la variation est juste dans l'architecture des différents niveaux du VAE, mais le fonctionnement et l'optimisation reste exactement la même comme présenté précédemment.

Annexe D

Présentation des outils informatiques employés

D.1 Présentation de Python

Python est un langage de programmation multi-paradigme (utilisé de plusieurs manières) moderne, interprété et fonctionne sur plusieurs systèmes (Windows, Mac, Linux). Le langage est principalement utilisé pour le calcul scientifique, la science des données, le développement web et comme un langage de scripting pour développer des extensions d'autres logiciels. Il comporte une très grande communauté et un grand nombre de bibliothèques open source, ceci lui a permis de se classer comme le premier langage de programmation utilisé. (*TIOBE Index*, 2022)

D.2 Présentation des bibliothèques employées

Une bibliothèque en programmation est un ensemble de modules, de méthodes, de classes et d'outils qui offrent une extension au langage utilisé, la majorité des bibliothèques existantes en Python sont disponibles gratuitement en open source. Dans cette partie, nous allons présenter les bibliothèques employées dans ce projet.

D.2.1 Pandas et NumPy

Pandas est une bibliothèque Python qui permet la manipulation des jeu de données (datasets) tabulaires, elle permet de lire des données de plusieurs sources telles que les fichiers CSV (comma separated values), des fichiers excels, des bases de données, etc. La bibliothèques offre un ensemble de méthodes pour filtrer, extraire, transformer et effectuer plusieurs opérations sur les données. La bibliothèque est elle-même basée sur la bibliothèque NumPy, qui est une autre bibliothèque de calculs et manipulation matricielle écrite en C et utilisable en Python, ce qui donne de très hautes performances, parfois plus rapide que l'usage des fonctionnalités natives de Python.

D.2.2 SciPy

Scipy est une bibliothèque pour le calcul scientifique avec un module de statistiques offrant un large nombre de tests statistiques, des méthodes d'interpolations et d'extrapolations, des modèles statistiques paramétriques et non paramétriques et autres fonctions. Cette

bibliothèque est souvent utilisée par d'autres bibliothèques de Machine Learning pour faciliter les calculs vu que SciPy existe depuis plusieurs années, est optimisé et a une grande communauté.

D.2.3 PyCaret

PyCaret est une bibliothèque utilisée pour développer des systèmes automatiques de machine learning. La bibliothèque permet d'entraîner un grand nombre de modèles en classification ou en régression ou autres, de trouver les paramètres optimaux et de déployer les modèles en quelques lignes de codes. Elle est pratique pour tester plusieurs modèles rapidement et trouver le modèle le plus convenable et ses paramètres optimaux en une courte durée. La bibliothèque est elle-même basée sur une autre bibliothèque très populaire de Machine Learning : Scikit-Learn, cependant cette dernière requiert l'entraînement et l'optimisation de chaque modèle manuellement et n'offre pas de mécanismes d'automatisation.

D.2.4 SDV

Synthetic Data Vault (SDV) est un écosystème de bibliothèques de génération de données synthétiques qui permet aux utilisateurs de générer facilement des ensembles de données à tableau unique, à tableaux multiples et à séries temporelles ayant le même format et les mêmes propriétés statistiques que l'ensemble de données original. SDV se base sur plusieurs techniques de modélisation graphique probabiliste et de Deep Learning.

D.2.5 Streamlit

Streamlit est une bibliothèque Python open-source qui facilite la création et le partage des applications web personnalisées qui se basent généralement sur le Machine Learning et la science des données. Streamlit permet, en quelques minutes seulement, créer et déployer de puissantes applications de données.

D.3 Présentation du Logiciel Any Logic

Le logiciel Anylogic a été développé initialement par un groupe de chercheurs de l'université de Saint-Pétersbourg afin de modéliser des processus parallèles dans le cadre de projets de recherches au sein de la société Hewlett Packard (HP). Ce projet fut développé en logiciel commercial en 2000 et il est aujourd'hui utilisé par un très grand nombre de sociétés à travers le monde tel que : IBM, Intel, PwC, British Airways, Nike, McDonald's, EPFL, SNCF, etc. Le logiciel permet de simuler n'importe quel système complexe à travers les différentes approches discutées précédemment, de plus il comporte un langage de modélisation graphique ainsi que la possibilité de modéliser à travers le langage de programmation JAVA. Pour ce travail, nous utiliserons la version "Personal Learning Edition", qui est une version gratuite d'apprentissage de ce logiciel.

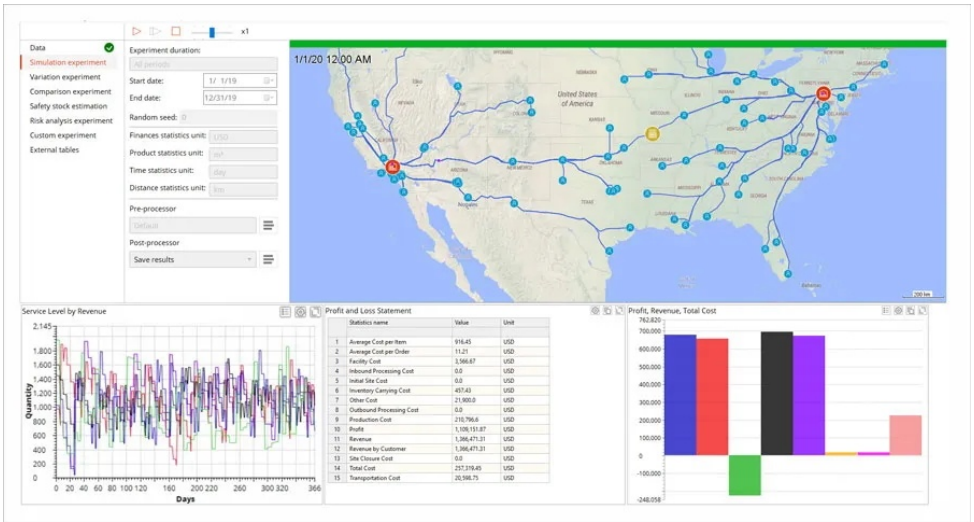
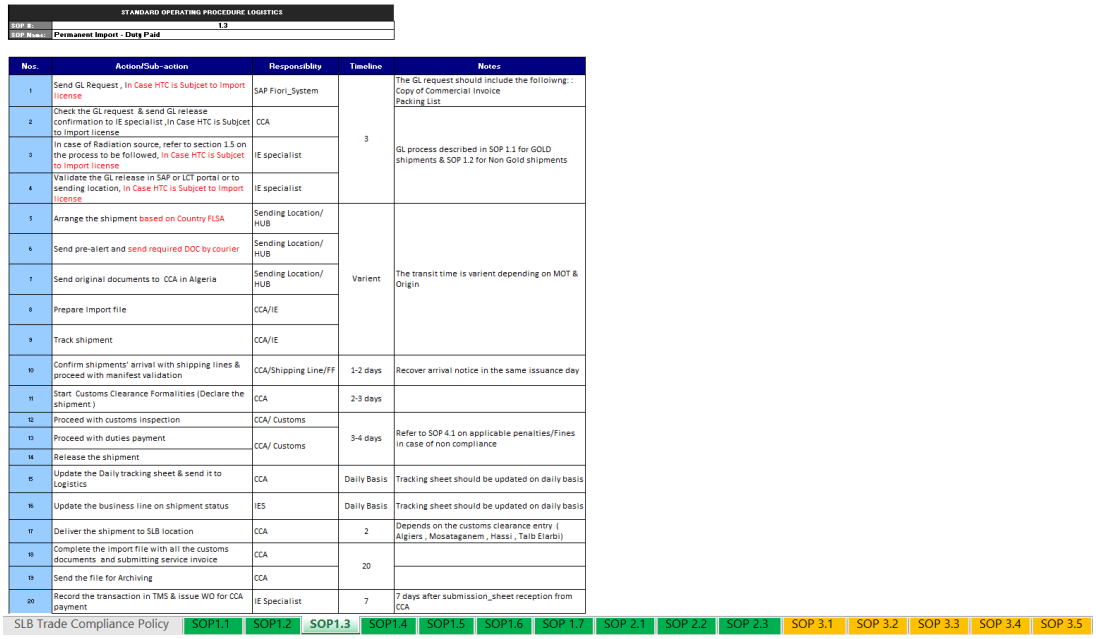


FIGURE D.1 – Aperçu du logiciel AnyLogic (AnyLogic, 2022)

Annexe E

Procédure standardisée du processus d'importation

C'est un document Excel qui englobe les différents cas possibles d'importation et détaille les différentes taches en précisant les responsabilités des parties prenantes et les échéanciers. Exemple : le cas d'une importation permanente avec paiement des droits et taxes.



STANDARD OPERATING PROCEDURE LOGISTICS	
SOP # :	1.3
SOP Name :	Permanent Import - Duty Paid

Nos.	Action/Sub-action	Responsiblity	Timeline	Notes
1	Send GL Request , In Case HTC is Subjcet to Import license	SAP Fiori_System	3	The GL request should include the folloiwng: : Copy of Commercial Invoice Packing List
2	Check the GL request & send GL release confirmation to IE specialist ,In Case HTC is Subjcet to Import license	CCA		GL process described in SOP 1.1 for GOLD shipments & SOP 1.2 for Non Gold shipments
3	In case of Radiation source, refer to section 1.5 on the process to be followed, In Case HTC is Subjcet to Import license	IE specialist		
4	Validate the GL release in SAP or LCT portal or to sending location, In Case HTC is Subjcet to Import license	IE specialist		
5	Arrange the shipment based on Country FLSA	Sending Location/ HUB	Varient	The transit time is varient depending on MOT & Origin
6	Send pre-alert and send required DOC by courier	Sending Location/ HUB		
7	Send original documents to CCA in Algeria	Sending Location/ HUB		
8	Prepare Import file	CCA/IE		
9	Track shipment	CCA/IE		
10	Confirm shipments' arrival with shipping lines & proceed with manifest validation	CCA/Shipping Line/FF	1-2 days	Recover arrival notice in the same issuance day
11	Start Customs Clearance Formalities (Declare the shipment)	CCA	2-3 days	
12	Proceed with customs inspection	CCA/ Customs	3-4 days	Refer to SOP 4.1 on applicable penalties/Fines in case of non compliance
13	Proceed with duties payment	CCA/ Customs		
14	Release the shipment			
15	Update the Daily tracking sheet & send it to Logistics	CCA	Daily Basis	Tracking sheet should be updated on daily basis
16	Update the business line on shipment status	IES	Daily Basis	Tracking sheet should be updated on daily basis
17	Deliver the shipment to SLB location	CCA	2	Depends on the customs clearance entry (Algiers , Mosataganem , Hassi , Talb Elarbi)
18	Complete the import file with all the customs documents and submitting service invoice	CCA	20	
19	Send the file for Archiving	CCA		
20	Record the transaction in TMS & issue WO for CCA payment	IE Specialist	7	7 days after submission_sheet reception from CCA

SLB Trade Compliance Policy | SOP1.1 | SOP1.2 | SOP1.3 | SOP1.4 | SOP1.5 | SOP1.6 | SOP 1.7 | SOP 2.1 | SOP 2.2 | SOP 2.3 | SOP 3.1 | SOP 3.2 | SOP 3.3 | SOP 3.4 | SOP 3.5

FIGURE E.1 – Exemple du processus standardisé d'importation permanente

Annexe F

Scorecard du contrat entre Schlumberger et les transitaires

Lead Time – Permanent (Working days)

KPI	MOT	Performance target	Penalty	Score
On Time Declaration form Complete file Date submission.	Courier	2 Days	**Hassi Messaoud**: 5% discount/per file for all the shipments released after the defined Lead Time **Other POE:** For any additional days after the defined Lead Time, Service provider will take all the logistics costs caused by the delay (storage, demurrage, customs penalties…)	**30%**
	Air Freight	2 Days		
	Sea freight	2 Days		
On Time Release from Customs **1- FD (Full duties)** **2- VAT exemption.**	Courier	3 Days – 5 Days		**20%**
	Air Freight	3 Days – 5 Days		
	Sea freight	3 days – 7 Days		
On Time Transport to Base	Courier	N/A		**10%**
	Air Freight	Form HMD to Base: 1 Day Form ALG to Base: 2 Days		
	Sea freight	Form ALG to Base: 2 Days Form MOST to Base: 3 Days		
On Time Container restitution	Sea freight	Form HMD to ALG: 2 Days		**10%**
On Time Invoicing		• 1 month for services • 1week for disbursment	5% credit note of the sum of invoices amounts deposited after the defined lead time	10%
Service Quality	Follow-up and closure of corrective actions	30 days after Schlumberger notification	5% credit note of the invoice related to the service involved	10%
	CCMS Service Quality incident	<3% of total monthly shipments (a)	5% credit note of the invoice related to the service involved	10%

FIGURE F.1 – Scorecard du transitaire

Annexe G

Code pour l'estimation du "Release lead time"

Cette annexe présente le code nécessaire pour développer et évaluer les modèles d'estimation du "Release leadtime", nous n'avons pas inclus les codes pour les autres lead times car l'approche est identique pour chaque lead time avec des petites modifications expliquées dans le chapitre 4, les explications de chaque partie sont en commentaire dans le code (délimitée par trois guillemets ou bien un symbole de dièse '')

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import jinja2
from pycaret.regression import *

"""## Chargement et traitement des données
Par la suite nous allons commencer par charger les données originales fournies par Schlumberger, on ne garde que les expéditions ayant un leadtime
totale infèrieur ou égale à 50 jours.
"""
original_df = pd.read_excel('clean_shipments.xlsx')
original_df = original_df[(original_df['Full delay'] >= 0) & (original_df['Full delay'] <= 50)]

"""De 196 expéditions,On se retrouve avec 186, donc 10 expéditions seulement qui ont été ignoré à travers le filtre précédent, ce qui est acceptable dans
 notre cas."""
original_df.shape
"""Par la suite nous allons charger les données synthétiques générées à travers la simulation. Nous allons séléctionner 500 expéditions générées"""
synth_sim_df = pd.read_excel('sim_output.xlsx')
synth_sim_df = synth_sim_df.sample(500,random_state=0)

"""Nous allons aussi charger les données synthétiques générée par le TVAE"""

synth_tvae_df = pd.read_excel('tvae_output.xlsx')
synth_tvae_df.head()

"""Avant de pouvoir fusionner les données originales et les données synthétiques de la simulation et du TVAE, il faudra s'assurer que les
 deux jeux de données ont les mêmes colonnes """
keep_cols = ['Status','Product Line','Legal Entity','Weight','Ship Unit Qty','Ship Unit',
             'Item Type', 'Port Of Entry', 'MOT', 'Regime','Service level',
             'Delivery Base', 'Customs Office', 'Customs Warehouse', 'CCA Name', 'Transporter',
             'Release leadtime'
             ]
missing_cols_sim = [col for col in keep_cols if col not in synth_sim_df.columns]
missing_cols_tvae = [col for col in keep_cols if col not in synth_tvae_df.columns]
print("Colonnes manquantes dans les données synth de simulation : ",missing_cols_sim)
print("Colonnes manquantes dans les données synth du TVAE : ",missing_cols_tvae)
"""
Toutes les colonnes de `keep_cols` sont présentes dans les données synthétiques et originales,
nous pouvons à présent les fusionner (nous allons aussi les permuter aléatoirement une fois fusionné) :
"""
```

```
df_1 = pd.concat([original_df[keep_cols],synth_sim_df[keep_cols]])
df_1 = df_1.sample(frac=1,random_state=10)
# DF 2 = Original data + TVAE data
df_2 = pd.concat([original_df[keep_cols],synth_tvae_df[keep_cols]])
df_2 = df_2.sample(frac=1,random_state=10)
"""Le nouveaux jeux de données `df_1` et `df_2` comporte 686 expéditions au lieu de 186 seulement. """
print(original_df['Release leadtime'].describe())
print(df_1['Release leadtime'].describe())
print(df_2['Release leadtime'].describe())

"""Nous pouvons à présent commencer l'entraînement des différents modèles de ML. Mais avant celà,
nous allons définir une méthode qui permet d'effectuer un encodage par la moyenne et d'appliquer une
 transformation logarithmique sur le poids."""
from pycaret.internal.utils import true_warm_start
from pandas.core.frame import DataFrame
from xlrd.formula import dump_formula

# Preprocess
target_col = 'Release leadtime'

def calc_smooth_mean(df, by, on, m):
    # Compute the global mean
    mean = df[on].mean()

    # Compute the number of values and the mean of each group
    agg = df.groupby(by)[on].agg(['count', 'mean'])
    counts = agg['count']
    means = agg['mean']

    # Compute the "smoothed" means
    smooth = (counts * means + m * mean) / (counts + m)

    # Replace each value by the according smoothed mean
    return df[by].map(smooth)

def preprocess(df,cols_me,w,keep_cols,log_weight=True):
  # Mean target encoding
  for col in cols_me:
    if col in keep_cols:
      df[col] = calc_smooth_mean(df,col,target_col,w)
  if log_weight:
    df['Weight'] = np.log1p(df['Weight'])
  return df
```

```
"""
Machine Learning
On commence pas définir les modèles qu'on utilisera à partir de la bibliothèque de PyCaret ainsi que les métriques calculés.
On définira aussi les colonnes aux-quels on appliquera l'encodage par la moyenne.
"""

include_models = ['rf','gbr','dummy','lr','dt','svm']
remove_metrics = ['MSE','MAPE','R2','RMSLE']
cols_to_encode = ['MOT', 'Product Line','Ship Unit','Item Type','Port Of Entry','Customs Office','Customs Warehouse',
                  'Delivery Base','Transporter'
                  ]

"""Nous allons calculer le MAE de base de référence d'un modèle "baseline" qui prévoit la moyenne à chaque fois."""

from sklearn.metrics import mean_absolute_error as mae
from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor
ybar = [original_df['Release leadtime'].mean()] * original_df.shape[0]
print(mae(original_df['Release leadtime'],ybar))

"""Cette érreur s'elève donc à 4.82 jours, on prendre le modèle qui donnera l'erreur la plus petite comparée à cette erreur de référence.

Avant de créer les modèles, nous allons mettre en place l'environnement de PyCaret,
nous obtenons une ensemble d'information sur le type de donnée dans chaque colonne. ainsi que les transformations effectué.
 Nous allons effectuer cela pour les 3 versions de données : les données originales `original_df,env_1`, les données originales fusionnées avec
 les données générées par la simulation `df_1,env_2`, et les données générées par le réseau de neurones TVAE `df_2,env_3`.

**Pour les données originales**
"""

# For original data

data = preprocess(original_df[keep_cols],cols_to_encode,20,keep_cols,True)
data['CCA_mean'] = calc_smooth_mean(data,'CCA Name','Release leadtime',10)
env_1 = setup(data,target='Release leadtime',session_id=2000,verbose=0,
              ordinal_features = {'Regime':['FD','VAT'],'Service level':['Standard','Urgent']},
              feature_selection = True, feature_selection_threshold = 0.6,
              numeric_features=['CCA_mean']
              )
rf = create_model(RandomForestRegressor(n_estimators=1,random_state=0),fold=4)
gbr = create_model(GradientBoostingRegressor(n_estimators=2,random_state=0),fold=4)
best_1 = compare_models(sort='MAE',include=[rf,gbr,'lr','svm','dt'],cross_validation=True,fold=4)
```

```
"""**Avec les données simulé**"""

data = preprocess(df_1[keep_cols],cols_to_encode,20,keep_cols)
data['CCA_mean'] = calc_smooth_mean(data,'CCA Name','Release leadtime',10)
env_2 = setup(data,target='Release leadtime',session_id=2000,verbose=0,
              ordinal_features = {'Regime':['FD','VAT'],'Service level':['Standard','Urgent']},
              feature_selection = True, feature_selection_threshold = 0.6,
              numeric_features=['CCA_mean']
              )
for metric in remove_metrics:
  remove_metric(metric)

best_2 = compare_models(sort='MAE',include=include_models,cross_validation=True,fold=4)

"""**Avec les données générées par le TVAE**"""

data = preprocess(df_2[keep_cols],cols_to_encode,20,keep_cols,True)
data['CCA_mean'] = calc_smooth_mean(data,'CCA Name','Release leadtime',10)
env_3 = setup(data,target='Release leadtime',session_id=2000,verbose=0,
              ordinal_features = {'Regime':['FD','VAT'],'Service level':['Standard','Urgent']},
              feature_selection = True, feature_selection_threshold = 0.6,
              numeric_features=['CCA_mean']
              )
for metric in remove_metrics:
  remove_metric(metric)
best_3 = compare_models(sort='MAE',include=['lr','gbr','rf','dt','svm'],n_select=4,cross_validation=True,fold=4)

"""Analysons le score d'apprentissage du meilleur modèle : """

plot_model(best_3[0],plot='learning')

"""
On remarque bien que le modèle peut bien réduire l'erreur globale sans tomber dans un cas d'overfitting, c'est donc un bon modèle à utiliser.
Les autres lead times sont estimé de la même manière que ce notebook.
"""
```