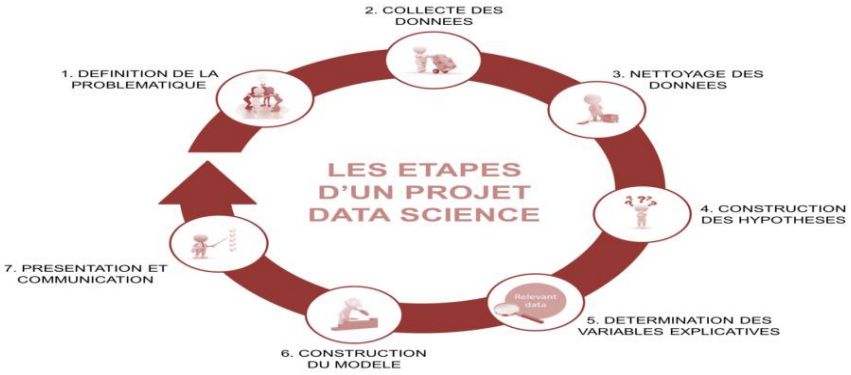
Chapitre II : Le processus de science des données



I. Introduction

La science des données continue d'évoluer comme l'un des parcours les plus prometteurs et les plus demandés pour les professionnels qualifiés. Aujourd'hui, les professionnels des données qui réussissent comprennent qu'ils doivent aller au-delà des compétences traditionnelles d'analyse de grandes quantités de données, d'exploration de données et de programmation. Afin de découvrir des renseignements utiles pour leurs organisations, les data scientists doivent maîtriser tout le spectre du cycle de vie de la science des données et posséder un niveau de flexibilité et de compréhension permettant de maximiser les rendements à chaque phase du processus.

II. Rôles et responsabilités dans un projet de science des données :

Les chercheurs d'informations viables peuvent distinguer les enquêtes importantes, rassembler des informations provenant d'un grand nombre de sources diverses, trier les données, interpréter les résultats dans des arrangements et transmettre leurs découvertes d'une manière qui influence décidément les choix des entreprises. Ces aptitudes sont requises dans pratiquement toutes les entreprises, ce qui fait que les chercheurs d'information doués sont progressivement importants pour les organisations.

Que fait un Data Scientist ?

Au cours de la dernière décennie, les chercheurs en information se sont révélés être des ressources fondamentales et sont disponibles dans pratiquement toutes les associations. Ces experts sont des personnes équilibrées, orientées vers l'information et dotées de compétences spécialisées de haut niveau, qui sont aptes à structurer des calculs quantitatifs complexes pour trier et orchestrer un grand nombre de données utilisées pour répondre aux questions et diriger la technique dans leur association. Ces compétences sont associées à une participation à la correspondance et à l'administration qui devrait permettre de transmettre des résultats substantiels à différents partenaires au sein d'une association ou d'une entreprise.

Les spécialistes des données doivent être curieux et axés sur les résultats, et posséder des connaissances exceptionnelles dans le domaine de l'industrie et des compétences en communication qui leur permettent d'expliquer des résultats hautement techniques à leurs homologues non techniques. Ils possèdent un solide bagage quantitatif en statistiques et en algèbre linéaire ainsi que des connaissances en programmation, notamment en matière d'entreposage de données, d'extraction et de modélisation pour construire et analyser des algorithmes.

Les algorithmes d'apprentissage automatique peuvent rassembler, stocker et analyser des données et générer un résultat valable. Ces outils vous permettent d'évaluer la situation à l'aide de données compliquées et groupées. On peut également dire que l'apprentissage automatique offre différents outils pour comprendre des données complexes par la segmentation et la simplification. En outre, il vous permet d'automatiser vos tâches professionnelles et de prendre de meilleures décisions grâce à des données organisées.

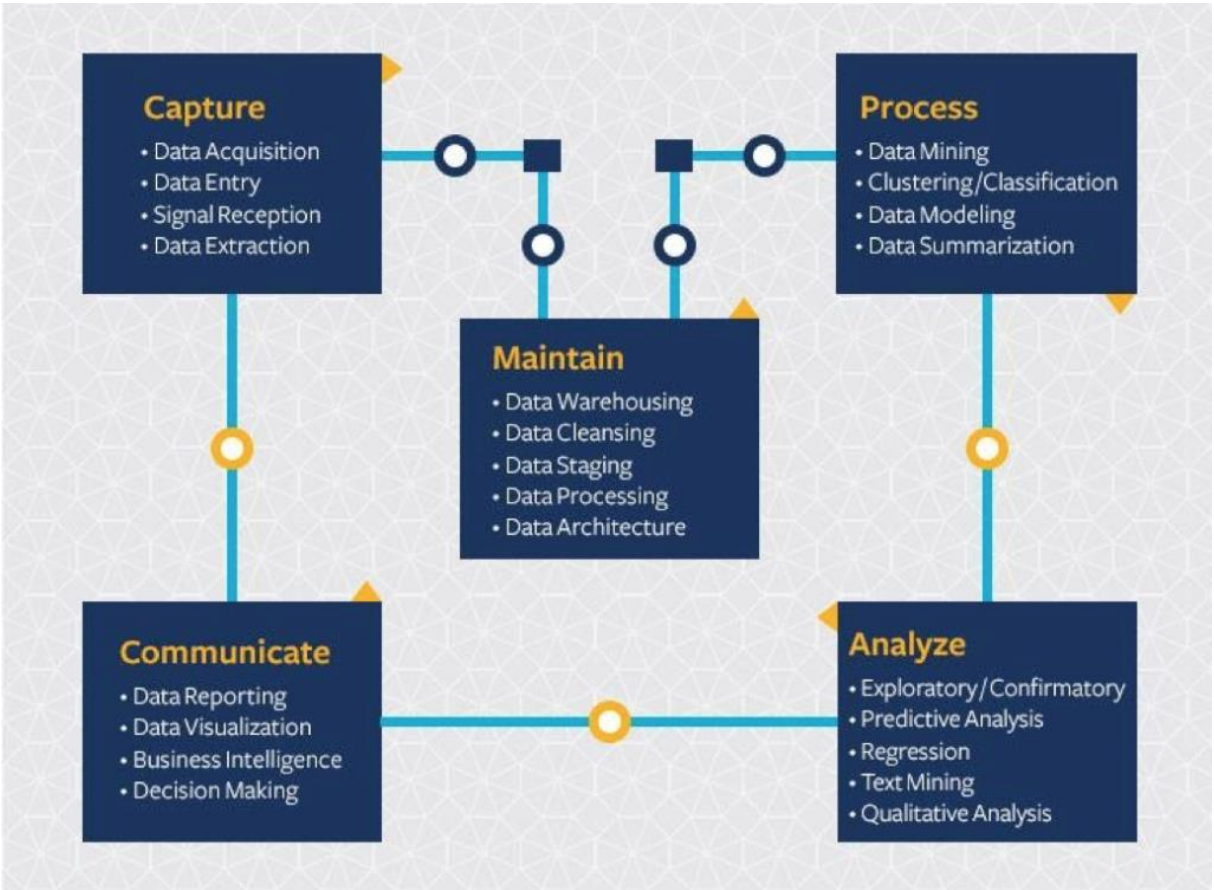
Certes, dans l'apprentissage automatique, les données servent de carburant. Vous introduisez de nouvelles données dans le modèle d'apprentissage automatique, et celui-ci génère le résultat souhaité en analysant toutes les données requises. L'algorithme utilisera des données pertinentes pour les résultats. Par conséquent, il est essentiel d'affiner les données de manière cohérente.

L'affinage permet de supprimer les données non pertinentes et obsolètes des ensembles de données. Vous n'avez plus besoin de ces données pour avoir un impact sur les résultats.

Les données non pertinentes dans un algorithme influenceront le résultat et affecteront la précision et le taux de réussite du modèle. Par conséquent, la suppression des données non pertinentes est essentielle pour apporter de l'efficacité au résultat. Par conséquent, cela clarifie l'importance du nettoyage des données dans l'apprentissage automatique.

III. Présentation du cycle de vie d'un projet de science des données

Le **cycle de vie d'un projet de science des données** est une démarche méthodique qui vise à structurer et organiser le processus d'exploration et d'analyse des données pour en extraire des informations pertinentes et exploitables. La figure suivante représente un aperçu de ce cycle de vie:



La figure représente les cinq étapes du cycle de vie de la science des données : Saisir (acquisition de données, saisie de données, réception de signaux, extraction de données) ; entretenir (stockage de données, nettoyage de données, mise en scène de données, traitement de données, architecture de données) ; traiter (exploration de données, classification/classement de données, modélisation de données, résumé de données) ; analyser (analyse exploratoire/confirmatoire, analyse prédictive, régression, exploration de textes, analyse qualitative) ; communiquer (rapport de données, visualisation de données, intelligence économique, prise de décision).

Etape 1 : Définir les objectifs de recherche et créer une charte de projet

Dans cette première phase, il est essentiel de **comprendre le problème à résoudre** et de **définir les objectifs** précis du projet. Cette étape implique la collaboration entre les parties prenantes pour

s'assurer que les questions posées répondent à des besoins réels de l'entreprise ou du domaine concerné. La charte de projet documente les objectifs, les livrables, les contraintes, les ressources disponibles et les délais à respecter. C'est aussi à ce moment que l'on clarifie les hypothèses, les attentes en termes de résultats et les critères de succès.

- **Définir les objectifs de recherche**

Les objectifs de recherche sont les **questions spécifiques** auxquelles le projet de science des données va tenter de répondre. Ils doivent être alignés avec les besoins des parties prenantes et fournir une orientation claire pour les analyses à mener. Voici les principales étapes pour définir ces objectifs:

a. Comprendre le contexte et les besoins

Avant tout, il est important de comprendre le **contexte métier** ou scientifique du projet. Cela implique de discuter avec les parties prenantes (décideurs, managers, experts métiers) pour saisir leurs attentes et les enjeux du projet. Quelques questions clés à poser :

- Quel problème souhaitez-vous résoudre avec les données ?
- Quels sont les indicateurs clés de performance (KPI) ou les résultats souhaités ?
- Qui sont les utilisateurs finaux des résultats (managers, clients, chercheurs, etc.) ?

b. Formuler les objectifs de recherche

Une fois que vous avez une bonne compréhension du contexte, les objectifs de recherche doivent être formulés de manière **claire** et **précise**. Voici quelques exemples d'objectifs :

- Prédire les ventes d'un produit au cours des trois prochains mois.
- Segmenter les clients selon leur comportement d'achat pour cibler des campagnes marketing.
- Identifier les facteurs déterminants de l'attrition des employés dans une entreprise.

Les objectifs doivent être **mesurables**, **atteignables**, **pertinents**, et **limités dans le temps** (SMART).

c. Identifier les hypothèses et les contraintes

Il est important de **clarifier les hypothèses** sous-jacentes du projet. Par exemple, une hypothèse pourrait être que les données disponibles sont suffisantes pour effectuer des prédictions fiables. Il faut également identifier les **contraintes** potentielles, comme des limitations en termes de données, de ressources humaines ou de temps.

- **Créer une charte de projet en science des données**

La **charte de projet** est un document formel qui décrit les éléments clés du projet de science des données. Elle sert à aligner tous les acteurs autour des mêmes objectifs et à définir les grandes lignes du projet. Voici les éléments principaux qui doivent y figurer :

a. Objectifs du projet

Cette section récapitule les objectifs définis précédemment, en précisant ce que le projet cherche à accomplir et pourquoi il est important. Il peut s'agir de questions exploratoires (comme identifier des patterns dans les données) ou de questions prédictives (comme prévoir des résultats futurs).

b. Périmètre du projet

Le périmètre (ou scope) définit les **limites** du projet : quelles données seront analysées, quelles méthodes seront utilisées, et quels résultats ou livrables sont attendus. Cela permet de gérer les attentes et d'éviter toute dérive du projet (scope creep).

c. Livrables attendus

Quels seront les **produits finaux** du projet ? Il peut s'agir de :

- Rapports d'analyse,
- Tableaux de bord interactifs,
- Modèles prédictifs déployés,
- Recommandations basées sur les résultats.

Cette section doit préciser les dates de livraison et les formats des livrables.

d. Données sources

Il est important d'identifier les **sources de données** qui seront utilisées dans le projet. Il peut s'agir de données internes ou externes (bases de données publiques, API). Cette section doit aussi mentionner si des étapes de collecte de nouvelles données sont nécessaires.

e. Plan de gestion des ressources

La charte de projet doit inclure les **ressources disponibles** pour le projet :

- Équipe projet : Qui sont les membres clés, leurs rôles (data scientist, analyste, expert métier, etc.) ?
- Technologies et outils : Quels outils de science des données seront utilisés (Python, R, SQL, etc.) ?
- Temps et budget : Quelles sont les ressources temporelles et financières affectées ?

f. Calendrier et jalons du projet

Le projet doit être découpé en **jalons** avec des échéances claires. Par exemple, un jalon pourrait être la fin de l'analyse exploratoire des données ou la validation d'un modèle prédictif. Cela permet de suivre l'avancement du projet.

g. Risques et gestion des imprévus

Tout projet de science des données comporte des **risques** (manque de données, faible qualité des données, difficulté à obtenir des résultats exploitables, etc.). La charte doit identifier ces risques et prévoir des **plans d'atténuation** (plan B).

h. Critères de succès

Enfin, il est essentiel de définir des **critères de réussite** du projet. Ces critères peuvent inclure des performances quantitatives (comme un taux de précision pour un modèle prédictif) ou des objectifs plus qualitatifs (satisfaction des parties prenantes).

Étape 2 : Récupération des données

La collecte et la gestion des données sont des étapes cruciales de la data science. Tout d'abord, les données peuvent provenir de différentes sources telles que les réseaux sociaux, les capteurs, les transactions bancaires, les systèmes de surveillance, etc.

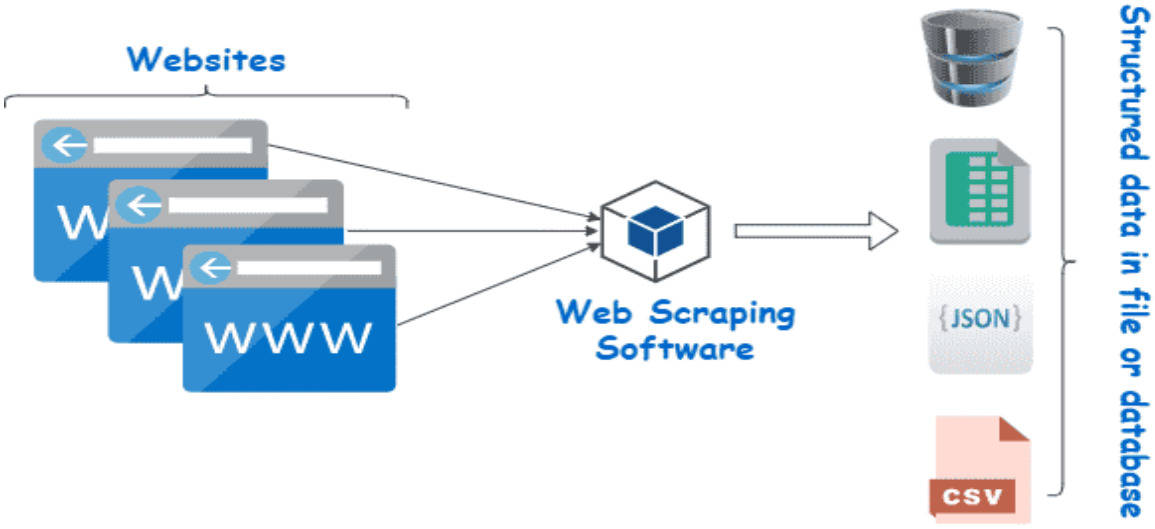
La collecte de données peut être effectuée de différentes manières, en fonction de la source de données et des besoins spécifiques. Pour stocker les données, il est important d'utiliser des techniques de stockage efficaces pour garantir que les données soient stockées de manière sécurisée et facilement accessibles.

Les outils de collecte

Les outils de gestion de données tels que les systèmes de gestion de bases de données (SGBD) permettent de stocker et d'organiser efficacement les données et de les rendre disponibles pour l'analyse. Les SGBD sont utilisés pour stocker une grande quantité de données et pour permettre un accès rapide et facile à ces données.

Il existe de nombreux outils de collecte de données disponibles pour collecter des données à partir de différentes sources. Voici quelques exemples d'outils de collecte de données populaires :

- **Web Scraping** : C'est une technique de collecte de données qui permet de récupérer des informations à partir de sites web. Les outils de web scraping populaires incluent BeautifulSoup, Scrapy, Selenium, etc.



- **API** : Les API (interfaces de programmation d'application) permettent de récupérer des données à partir de sources en ligne telles que les réseaux sociaux, les services de streaming, etc. Les API populaires incluent Twitter API, Facebook Graph API, Google Maps API, etc.
- **Capteurs** : Les capteurs peuvent être utilisés pour collecter des données à partir d'appareils électroniques tels que les smartphones, les montres intelligentes, les appareils de santé, etc.
- **Enquêtes et sondage**s : Les enquêtes et les sondages sont une méthode populaire pour collecter des données à partir de personnes. Les outils de sondage en ligne populaires incluent SurveyMonkey, Google Forms, Typeform, etc.

- **Données publiques** : Les données publiques sont des données rendues disponibles par les gouvernements ou les organisations publiques. Les sources de données publiques populaires incluent le site web de l'OMS, le site web de la Banque mondiale, etc.

Lorsque l'on commence à s'intéresser à la data science, les **données publiques sont parfaites pour s'entrainer au traitement de données**. Il existe de nombreux sites qui fournissent des données publiques pour différentes régions du monde et différents domaines tels que la santé, l'environnement, l'économie, l'éducation, la démographie, etc. Les données publiques sont des données rendues disponibles par les gouvernements ou les organisations publiques. Elles peuvent être utilisées pour la recherche, l'analyse et la prise de décision. Voici quelques exemples de sites pour accéder à des données publiques :

- **Data.gov** : une plateforme de données ouverte du gouvernement américain qui fournit un accès à plus de 200 000 ensembles de données. L'équivalent en France est sur data.gouv.fr.
- **Eurostat** : la base de données statistiques de l'Union européenne qui fournit des données sur l'économie, la société et l'environnement en Europe.
- **UNdata** : la base de données statistiques des Nations unies qui fournit des données sur les pays membres des Nations unies dans de nombreux domaines.
- **OECD** : l'Organisation de coopération et de développement économiques fournit des données sur les pays membres et les économies émergentes.
- **World Bank Open Data** : une plateforme de données ouverte de la Banque mondiale qui fournit des données sur le développement économique et social dans le monde entier.
- **US Census Bureau** : l'agence statistique du gouvernement américain qui fournit des données démographiques, économiques et sociales sur les États-Unis.

Il est important de choisir les sources de données en fonction de leur fiabilité et de leur pertinence pour le projet en question. Les données publiques sont un excellent moyen d'accéder à des informations précieuses sur divers sujets et de faciliter l'analyse de données à grande échelle.

Étape 3 : Nettoyer, intégrer et transformer les données

Avant de commencer toute analyse, il est indispensable de **préparer les données**. Cela inclut :

- Le **nettoyage des données** pour supprimer ou corriger les erreurs (valeurs manquantes, doublons, outliers),
- L'**intégration des données** provenant de différentes sources, afin de les rendre compatibles et utilisables conjointement,
- La **transformation des données**, c'est-à-dire la normalisation, l'encodage des variables catégorielles, l'agrégation, ou la création de nouvelles variables si nécessaire.

Cette étape est cruciale, car des données de mauvaise qualité peuvent conduire à des résultats biaisés ou incorrects.

La première étape du nettoyage des données consiste à identifier vos objectifs. Vous ne pouvez pas accomplir vos tâches si vous n'avez aucune idée de vos attentes. Une fois que vous connaissez vos objectifs, vous pouvez mettre en place un plan pour les atteindre. Dans ce cas, votre objectif principal est d'apporter de la précision et de supprimer les erreurs. Pendant la planification, vous choisirez la stratégie à suivre. Commencer par se concentrer sur les principaux paramètres serait la meilleure décision. Cependant, vous devez vous poser quelques questions afin de trouver les bons indicateurs.

- Quelle serait la métrique la plus élevée pour atteindre le résultat souhaité ?
- Quelles sont vos attentes en matière de nettoyage des données ?

Une fois que vous avez compris la raison pour laquelle vous devez nettoyer les données, vous pouvez suivre les étapes suivantes :

- **Identifier les erreurs**

Avant de corriger l'erreur et d'apporter de la précision à la sortie du modèle, vous devez d'abord l'identifier. L'identification des erreurs vous aidera à trouver la solution optimale en un minimum de temps. Cependant, l'évaluation de données complètes peut être intimidante et peut affecter les fonctions des modèles. Conservez donc un registre de tous les ensembles de données où vous rencontrez le plus d'erreurs. La tenue de ces registres vous permet de simplifier le processus d'identification et de résolution des données corrompues ou incorrectes.

- **Normaliser le processus**

Tout en nettoyant les données, vous devez également reconnaître si l'erreur est due à une valeur incorrecte. Chaque valeur de données doit être dans un format standardisé. Par exemple, vous devez vérifier les minuscules et les majuscules des chaînes de caractères ou mesurer l'unité des valeurs numériques. Il arrive que le modèle considère les données comme inexactes en raison de telles coquilles et erreurs.

- **Vérifiez l'exactitude des données**

Après avoir analysé la base de données pour le nettoyage des données, confirmez l'exactitude des données à l'aide de différents outils. Vous devez investir dans des outils de données pour rationaliser et accélérer le processus de nettoyage. La plupart de ces outils utilisent un algorithme d'apprentissage automatique pour identifier les données appropriées et les nettoyer en temps réel. Par la suite, cela a un impact positif sur la précision du modèle et génère les meilleurs résultats.

- **Vérifiez les données en double**

Les données en double peuvent ne pas causer d'erreur, mais elles font perdre beaucoup de temps au résultat. Cependant, vous pouvez résoudre ce problème en identifiant les doublons pendant l'analyse des données. Recherchez des outils d'analyse de données pour nettoyer les données des doublons. Choisissez un outil automatisé pour analyser et supprimer les données en double.

L'importance du nettoyage des données

Comme dans beaucoup d'entreprises, les données peuvent être d'une importance capitale pour votre entreprise. Avec des données précises, vous pouvez améliorer vos opérations commerciales et prendre de meilleures décisions. Par exemple, vous êtes une entreprise de livraison, et votre activité dépend de l'adresse de vos clients. Pour que les données restent exactes, vous devez constamment mettre à jour la base de données. Comme de nombreux clients de la ville peuvent changer de quartier, vous devez mettre à jour les données régulièrement. Si vos données sont inexactes et périmées, vos employés

commettront des erreurs lors de l'exécution de leurs tâches professionnelles. Par conséquent, concentrez-vous sur la mise à jour des nouvelles données et le nettoyage des anciennes. Voici quelques avantages du nettoyage des données pour votre entreprise :

- ✓ Technique rentable
- ✓ Réduit les risques d'erreurs
- ✓ Améliore l'acquisition de clients
- ✓ Augmentation des données homogènes
- ✓ Prise d' une meilleure décision
- ✓ Augmentation de la productivité des employés

Étape 4 : Analyse exploratoire des données

L'analyse exploratoire (ou **EDA** pour "Exploratory Data Analysis") permet de **mieux comprendre les données**. Les data scientists utilisent des techniques statistiques et des visualisations (graphiques, histogrammes, nuages de points, etc.) pour :

- Identifier les **tendances** et **patterns** dans les données,
- Détecter les **corrélations** entre les variables,
- Identifier les **anomalies** ou les données aberrantes,
- Mieux comprendre la distribution et la structure des données.

Cette phase est souvent itérative et aide à préparer la modélisation.

L'EDA permet de déterminer la meilleure façon de manipuler les sources de données pour obtenir les réponses recherchées. Elle permet aux data scientists de plus facilement découvrir des schémas, repérer des anomalies, tester des hypothèses ou vérifier des suppositions.

L'EDA est principalement utilisée pour découvrir ce que les données peuvent révéler au-delà de la modélisation formelle ou du test d'hypothèses, et elle permet de mieux comprendre les variables des jeux de données et les relations entre elles. Elle permet également de déterminer si les techniques statistiques que vous envisagez d'utiliser pour l'analyse des données sont adaptées. Développées à l'origine par le mathématicien américain John Tukey dans les années 1970, les techniques d'EDA restent aujourd'hui une méthode largement utilisée dans le processus de découverte de données.

Pourquoi l'analyse exploratoire des données est-elle importante dans le domaine de la science des données ?

L'objectif principal de l'EDA, c'est de vous aider à examiner les données avant de faire des suppositions. Elle peut permettre d'identifier les erreurs évidentes, de mieux comprendre les schémas dans les données, de détecter les données aberrantes ou les événements anormaux, et de trouver des relations intéressantes entre les variables.

Les data scientists peuvent utiliser l'analyse exploratoire pour s'assurer que les résultats produits sont valides et applicables à tous les résultats commerciaux et objectifs métier visés. L'EDA permet également aux parties prenantes de confirmer qu'elles posent les bonnes questions. L'EDA peut vous aider à répondre aux questions que vous avez sur les écarts-types, les variables nominales et les intervalles de confiance. Une fois l'EDA terminée et les informations déduites, ses fonctionnalités

peuvent être utilisées pour une analyse ou une modélisation des données plus sophistiquée, y compris le machine learning.

Outils d'analyse exploratoire des données

Les fonctions et techniques statistiques spécifiques que vous pouvez exécuter avec les outils EDA sont notamment les suivantes :

- Les techniques de clustering et de réduction de la dimensionnalité, qui permettent de créer des représentations graphiques des données de grande dimension comptant de nombreuses variables.
- La visualisation univariée de chaque champ du jeu de données brutes, avec des statistiques récapitulatives.
- Des visualisations bivariées et des statistiques récapitulatives qui vous permettent d'évaluer la relation entre chaque variable du jeu de données et la variable cible examinée.
- Visualisations multivariées, pour mapper et comprendre les interactions entre les différents champs des données.
- Le partitionnement en k-moyennes ou k-means clustering est une méthode utilisée dans l'apprentissage non supervisé où les points de données sont divisés en k groupes, c'est-à-dire le nombre de clusters, en fonction de la distance par rapport au centroïde de chaque groupe. Les points de données les plus proches d'un centroïde particulier seront regroupés dans la même catégorie. Le clustering en k-moyennes est couramment utilisé dans la reconnaissance de formes et la compression d'images.
- Les modèles prédictifs, tels que la régression linéaire, utilisent des statistiques et des données pour prédire des résultats.

Étape 5 : Construire les modèles

Après l'analyse exploratoire, l'étape suivante consiste à **développer des modèles prédictifs ou descriptifs** à partir des données. Cela peut inclure des modèles de **machine learning** (régression, classification, clustering, etc.) ou d'autres types d'algorithmes. Les étapes clés ici incluent :

- La **sélection des modèles**,
- Le **test et l'entraînement des modèles** sur des jeux de données d'entraînement,
- L'évaluation des performances des modèles à l'aide de métriques appropriées (précision, rappel, RMSE, etc.).

Cette étape est également souvent itérative, avec des ajustements constants pour améliorer la performance des modèles.

Voici un aperçu détaillé des étapes et des meilleures pratiques pour construire les modèles dans un projet de science des données.

1. Préparation des données pour la modélisation

Avant de commencer à créer les modèles, il est souvent nécessaire de **préparer les données** de manière spécifique. Cela peut inclure plusieurs opérations, comme :

a. Division des données en ensembles d'entraînement et de test

- L'ensemble **d'entraînement** est utilisé pour **entraîner** le modèle (souvent environ 70-80 % des données).
- L'ensemble **de test** est utilisé pour **évaluer les performances** du modèle une fois qu'il a été entraîné (environ 20-30 % des données).
- Parfois, un troisième ensemble, appelé **ensemble de validation**, est utilisé pour ajuster les hyperparamètres des modèles.

b. Normalisation ou standardisation des données

Certains algorithmes (comme la régression linéaire ou les réseaux de neurones) nécessitent que les données soient **normalisées** ou **standardisées**. Cela permet d'éviter que des variables avec des échelles différentes influencent disproportionnellement le modèle.

- **Normalisation** : transformation des données pour qu'elles aient des valeurs entre 0 et 1.
- **Standardisation** : centrage des données autour de la moyenne 0 avec un écart-type de 1.

c. Encodage des variables catégorielles

Si votre jeu de données comporte des variables catégorielles (par exemple, des valeurs non numériques comme « genre », « type de produit »), elles doivent être **encodées** sous forme numérique pour être comprises par les algorithmes. Deux approches courantes :

- **Encodage one-hot** : crée une colonne binaire (0 ou 1) pour chaque catégorie.
- **Encodage ordinal** : attribue un nombre entier à chaque catégorie.

d. Gestion des déséquilibres dans les classes

Si vous avez des classes déséquilibrées (par exemple, si dans un problème de classification, 90 % des échantillons appartiennent à une classe et 10 % à une autre), vous devrez peut-être utiliser des techniques pour **rééquilibrer les classes** :

- **Sous-échantillonnage** de la classe majoritaire,
- **Suréchantillonnage** de la classe minoritaire,
- Utilisation de la méthode **SMOTE** (Synthetic Minority Over-sampling Technique) pour générer des échantillons synthétiques de la classe minoritaire.

2. Choix des algorithmes de modélisation

Le choix du ou des algorithmes de modélisation dépend de la nature du problème à résoudre (classification, régression, clustering, etc.) et des caractéristiques des données. Voici les principaux types de modèles utilisés :

a. Modèles supervisés

Ces modèles sont utilisés lorsqu'on dispose de données étiquetées (où la variable cible est connue). Ils servent principalement pour les tâches de classification et de régression.

- **Régression linéaire** : utile pour des problèmes de **régression** (prédiction d'une valeur numérique) comme la prédiction des ventes ou des prix.

- **Régression logistique** : pour les problèmes de **classification binaire** (oui/non, succès/échec, etc.).
- **Arbres de décision et forêts aléatoires (random forests)** : permettent de gérer des problèmes de classification ou de régression avec des variables complexes, en utilisant des structures arborescentes.
- **Support Vector Machines (SVM)** : pour des problèmes de classification, surtout lorsque les données sont bien séparables.
- **Réseaux de neurones artificiels (ANN)** : utiles pour des problèmes complexes comme les images ou le traitement du langage naturel (plus souvent utilisés dans des approches de deep learning).

b. Modèles non supervisés

Ces modèles sont utilisés pour des données sans étiquette (pas de variable cible). Ils sont souvent employés pour découvrir des motifs ou des structures sous-jacentes dans les données.

- **K-means clustering** : pour regrouper des données similaires en clusters.
- **Algorithme DBSCAN** : pour le clustering avec des formes complexes et des données bruitées.
- **ACP (Analyse en Composantes Principales)** : pour la **réduction de dimension** et la simplification des jeux de données complexes.

c. Modèles semi-supervisés ou apprentissage par renforcement

- **Apprentissage semi-supervisé** : combinaison de données étiquetées et non étiquetées pour les tâches de classification.
- **Apprentissage par renforcement** : utile dans des situations où un agent apprend par essais et erreurs, en interaction avec un environnement dynamique (exemple : apprentissage de jeux, robots autonomes).

3. Entraînement des modèles

L'entraînement d'un modèle consiste à **ajuster les paramètres du modèle** en fonction des données d'entraînement pour qu'il soit capable de faire des prédictions précises sur de nouvelles données. Voici comment cela se passe :

a. Sélection d'un algorithme

En fonction du type de problème (classification, régression, etc.) et des caractéristiques des données, vous choisissez un ou plusieurs algorithmes pour commencer l'entraînement.

b. Ajustement des hyperparamètres

Chaque algorithme possède des **hyperparamètres**, qui sont des paramètres externes au modèle et qu'il faut ajuster manuellement pour optimiser les performances (par exemple, la profondeur d'un arbre de décision, le taux d'apprentissage d'un réseau de neurones). Des techniques comme la **recherche par grille** (grid search) ou la **recherche aléatoire** (random search) sont souvent utilisées pour trouver les meilleurs hyperparamètres.

c. Validation croisée

La **validation croisée** est une technique pour éviter le surapprentissage (overfitting). Elle consiste à diviser les données en plusieurs sous-ensembles (folds), à entraîner le modèle sur certaines parties des données et à le tester sur d'autres. Cela permet d'évaluer la capacité de généralisation du modèle.

4. Évaluation des modèles

Une fois le modèle entraîné, il doit être **évalué** pour s'assurer qu'il répond aux objectifs du projet. Les performances d'un modèle sont mesurées à l'aide de métriques spécifiques selon le type de problème:

a. Métriques de classification

- **Précision (Accuracy)** : proportion de prédictions correctes parmi toutes les prédictions effectuées.
- **Précision, rappel et F1-score** : pour évaluer la qualité des prédictions positives, particulièrement utile dans des cas de déséquilibre des classes.
- **Matrice de confusion** : tableau récapitulatif des prédictions correctes et incorrectes.

b. Métriques de régression

- **Erreur quadratique moyenne (MSE)** ou **racine de l'erreur quadratique moyenne (RMSE)** : mesure la différence moyenne au carré entre les valeurs prédites et les valeurs réelles.
- **R^2 (coefficient de détermination)** : indique la proportion de variance expliquée par le modèle.

5. Amélioration et optimisation des modèles

Après l'évaluation initiale, il est souvent nécessaire de **réajuster** ou d'**optimiser les modèles** pour améliorer leurs performances.

a. Réglage des hyperparamètres

Comme mentionné précédemment, des techniques de réglage des hyperparamètres peuvent être appliquées pour améliorer les performances globales du modèle.

b. Ensembles de modèles

Parfois, la combinaison de plusieurs modèles (techniques d'**ensemble**) peut offrir de meilleurs résultats

- **Bagging** : comme les forêts aléatoires, combine plusieurs modèles pour réduire la variance.
- **Boosting** : algorithme d'ensemble qui combine plusieurs modèles faibles pour en créer un plus fort (comme XGBoost ou AdaBoost).

c. Surapprentissage (Overfitting)

Si le modèle est trop performant sur les données d'entraînement mais échoue sur les données de test, il est probable qu'il ait surappris. Des techniques comme la **régularisation** peut être utilisée.

Construire des modèles est une étape complexe et critique dans un projet de science des données. Elle implique non seulement de choisir les bons algorithmes en fonction des données et du problème, mais aussi de s'assurer que le modèle est bien entraîné, correctement évalué et ajusté pour offrir les meilleures performances possibles dans le cadre des objectifs du projet.

Étape 6 : Présentation des résultats et création d'applications

Une fois que les modèles ont produit des résultats satisfaisants, il est essentiel de **communiquer ces résultats** aux parties prenantes. Cela peut se faire sous forme de rapports, de **tableaux de bord interactifs**, ou d'applications web qui permettent de visualiser les prédictions ou les insights extraits des données. Parfois, des **applications de machine learning** ou des solutions automatisées peuvent être construites à partir des modèles, pour être intégrées dans des processus opérationnels. Il est important que les résultats soient compréhensibles et exploitables par les utilisateurs finaux.

Ce cycle de vie n'est pas strictement linéaire : il peut y avoir des allers-retours entre les étapes. Par exemple, de nouvelles découvertes lors de l'analyse exploratoire peuvent amener à ajuster la collecte de données ou à transformer les données différemment.