Introduction au data Science

Cours réalisé Par GHERARI MANEL

Master 1 GADM

Année 2021-2022

Chapitre 1 : Introduction au Data Science

1. Table des matières

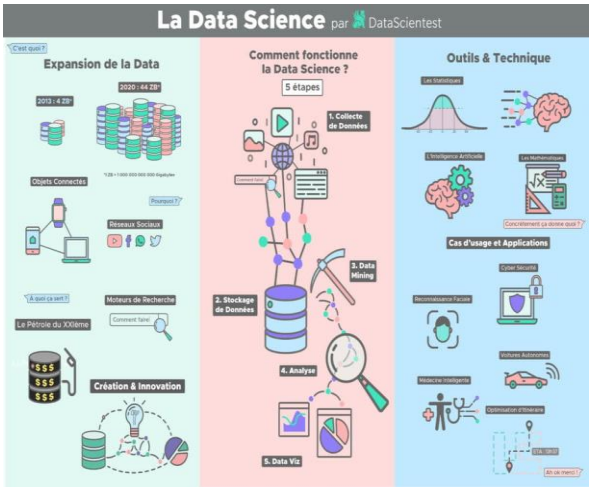
2. Introduction 4
3. Data Science- Definition 4
4. Pourquoi la science des données ? 4
5. Difference entre Data Scientist, data analyst data engineer 5
Data Analyst 6
Business Intelligence Developer 6
Data Engineer 7
Data Scientist 8

6. Comment fonctionne la data science ? 8
6.1. Collecte des données : Recueillir les données pertinentes du domaine 9
6.2. Trouver les bons outils pour chaque tâche 10
6.3. Préparation des données : Extraire et formater 11
6.4. Réduire la dimensionnalité des données 12
6.5. Transformer les données pour répondre aux exigences des algorithmes 13
6.6. Analyse des données : Conduire les analyses avec une modélisation efficace 15
7. Les cas d'usage et applications 18

1. Introduction

La Data Science ou science des données est un vaste champ multi-disciplinaire visant à donner du sens aux données brutes. Data Science : définition, champs d'applications et limites actuelles, découvrez tout ce que vous devez savoir sur ce domaine complexe, devenu un enjeu prioritaire dans les entreprises de toutes les industries.

2. Data Science- Definition



Pour définir la Data Science de la plus simple des façons, il s'agit de l'extraction d'informations exploitables à partir de données brutes. Ce champ multi-disciplinaire a pour but principal d'identifier des tendances, des motifs, des connexions et des corrélations dans les larges ensembles de données. La science des données englobe une large variété d'outils et de techniques telles que la programmation informatique, l'analyse prédictive, les mathématiques, les statistiques ou l'intelligence artificielle. Désormais, la Data Science inclut aussi les algorithmes de Machine Learning. De nos jours, presque toutes les entreprises affirment pratiquer la Data Science sous une forme ou une autre. Cependant, les méthodes et approches employées peuvent varier d'une organisation à l'autre. Il devient donc très compliqué d'offrir une définition précise de la Data Science. D'autant que de nouvelles technologies apparaissent sans cesse et transforment continuellement ce domaine. Ainsi, pour définir la science des données, la meilleure question à se poser est " pourquoi ? ".

3. Pourquoi la science des données ?

Si la Data Science connaît un essor fulgurant dans tous les secteurs d'activité, c'est parce que l'humanité génère de plus en plus de données. Entre 2011 et 2013, en seulement deux ans, le volume mondial de données a été multiplié par 9.

Et cette explosion du Big Data n'a pas ralenti depuis. D'ici la fin de l'année 2020, le volume total de données à l'échelle de la planète devrait atteindre 44 zettabytes contre moins de 5 zettabytes en 2013.

Comment expliquer ce phénomène ? Plusieurs technologies émergentes génèrent des données. C'est le cas des objets connectés, des réseaux sociaux, des smartphones, ou des moteurs de recherche web.

Or, toutes ces données offrent des opportunités inouïes pour les entreprises de toutes les industries, les institutions de recherche ou le secteur public. C'est la raison pour laquelle les données sont souvent considérées comme " le pétrole du XXIème siècle ".

Le but de la Data Science est d'exploiter ces données, de leur donner du sens. Cette discipline vise à parcourir de vastes " lacs de données " à la ***recherche de connexions, de tendances, de points d'intérêt****.*

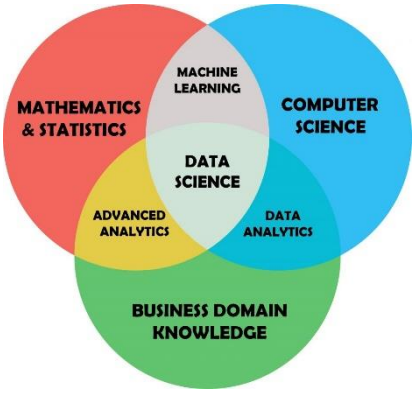
En s'appuyant sur ces découvertes, il est possible de créer de nouveaux produits et services innovants, de résoudre des problèmes concrets, d'améliorer ses performances comme jamais auparavant. La Data Science permet de prendre des décisions basées sur les données, plutôt que sur une simple intuition. Ainsi, elle révolutionne notre quotidien et nous permet de s'ouvrir à de nouveaux horizons. En bref, la data science représentera une science incontournable du monde demain !

4. Difference entre Data Scientist, data analyst data engineer

En science des données, résoudre des problèmes et répondre à des questions par l'analyse est une pratique standard. Les spécialistes des données expérimentent en continu en construisant des modèles pour prédire des résultats ou découvrir des schémas sous-jacents, avec pour objectif d'obtenir de nouveaux éclairages. Mais ils ne peuvent réaliser cela sans support.

C'est là que l'ingénieur des données entre en jeu. Avec les responsables métier, les responsables informatiques et d'autres professionnels compétents, les spécialistes et les ingénieurs des données constituent une équipe élargie dans laquelle chaque membre joue un rôle essentiel dans l'exploitation des données internes et externes, afin de donner un avantage concurrentiel à l'entreprise et prendre de meilleures décisions.

À mesure que l'analyse de données devient plus accessible et se répand, de nouvelles pratiques émergent – dont l'utilisation de très grands ensembles de données, l'intégration de l'analyse de texte dans les modèles prédictifs, ou l'automatisation accrue. Pour aider à guider les processus et les activités d'un secteur, les spécialistes et les ingénieurs des données doivent adopter la vue centrée sur les données d'une méthodologie fondamentale de la science des données.



Ce livret décrit les composantes et les processus qui constituent la méthodologie fondamentale de la science des données, et aborde certains des outils et techniques aujourd'hui utilisés par les ingénieurs des données pour collecter, traiter, analyser et déployer les données. Mais avant aller plus loin, il fauts distinguée entre les 3 acteurs prinipale dans le domane de data science (i.e, Data analyst, Data engineer, data Scientist).

Data Analyst

Un analyste de données, est un quelqu'un qui est capable d'interroger des sources de données pour en faire des rapports et des visualisations graphiques (graphes camemberts, histogrammes etc…). Un Data Analyst a une compréhension forte du domaine métier dans lequel il opère. Ce qui lui permet de mieux communiquer avec les gens du métier.

Pour mieux explorer les données, un Data Analyst est généralement à l'aise avec les outils statistiques. Toutefois, il n'est pas forcement aussi "calé" techniquement qu'un software engineer pour traiter les grands volumes de données (Big Data).

Compétences et outils : Excel, Access, SQL, SPSS, Tableau, Statistiques.

Business Intelligence Developer

Les développeurs B.I. (Business Intelligence / informatique décisionnelle) vont mettre en place des outils de B.I. pour les besoins de l'entreprise. Ces outils se présentent généralement sous forme de *Data warehouses, Datamart,* ainsi que des *bases de données multidimensionnelles* construits à partir **d'agrégation de données** en provenance de plusieurs bases de données. La construction des *Data warehouse et les bases OLAP* est généralement

effectuée à travers des Job ETL (Extract, Transform, Load) en utilisant l'outil **Talend** par exemple.

Ces Bases de données multidimensionnelles et *Data warehouses* sont par la suite utilisées par les développeurs B.I pour construire des tableaux de bords (*Dashboards)* et des rapports utiles pour les manageurs et les décideurs.

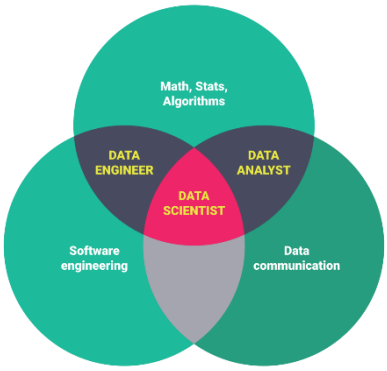
Les développeurs de B.I. ont généralement une connaissance métier moindre que celle d'un Data Analyst. Cependant, ils sont plus "calés" techniquement pour s'interfacer avec les différentes sources de données.

Compétences et outils : SQL, OLAP, Data warehouses, Cubes, SSAS, SSIS, ETL (Talend…)

Data Engineer

Un Data Engineer est quelqu'un ayant un background technique en développement logiciel. Il peut être un Software Engineer qui s'est reconverti dans le Big Data.

Les Data Engineers vont mettre en place des systèmes de Big Data pour traiter ces dernières. Ils opteront pour des outils de stockage performants comme les bases de données NoSQL et se baseront sur Hadoop, Spark, Map/Reduce pour traiter convenablement ces grands volumes de données.



Les Data Engineer vont collecter, transformer les données de différentes sources. Ce travail préparatoire permettra d'avoir des données "propres", prêtes pour qu'on leur applique dessus des techniques de Machine Learning.

En d'autres termes, le travail d'un Data Engineer est de préparer le terrain pour qu'un Data Scientist puisse se servir des données propres pour en tirer des tendances (Insights).

Compétences et outils : SQL, NoSQL, Hadoop, Data Lake, Big Data, Spark, Software Engineering, Map/Reduce…

Data Scientist

Un Data Scientist est un profil pluridisciplinaire qui aura pour mission première de tirer de l'information utile (insights) depuis des données brutes. Le métier du Data Scientist est à l'intersection entre Data Analyst et de Data Engineer. Tout en ayant des connaissances métiers dans le domaine dans lequel il évolue.

En effet, un Data scientist va explorer et exploiter les gisements de données de l'entreprise pour leur appliquer des techniques de machine learning. Il s'agit donc d'une forme de Data Analysis poussée sur de grands volumes de données. L'exposition au contexte Big Data exige qu'un Data Scientist soit familier avec des concepts comme Map/Reduce, Hadoop, Data lake etc…

L'information utile recherchée par un Data Scientist est spécifique à une entreprise et plus généralement à un domaine métier. Pour cela, un Data Scientist doit être à l'aise avec le domaine métier dans lequel il opère. Pour cela, il côtoiera les gens du métier pour creuser avec eux les différentes pistes de réflexion.

Finalement, un data scientist doit être un bon communicant pour mieux communiquer ses retrouvailles. Il usera pour cela des différents supports de présentation comme les présentations PowerPoint, ainsi que des visualisations graphiques (histogrammes, camemberts…) plus parlantes aux décideurs.

Compétences et outils : SQL, NoSQL, Python, R, Machine Learning, Deep Learning, Statistiques, Software Engineering…

5. Comment fonctionne la data science ?

La Data Science couvre une large variété de disciplines et de champs d'expertise. Son but reste toutefois de donner du sens aux données brutes. Pour y parvenir, les Data Scientists doivent posséder des compétences en ingénierie des données, en mathématiques, en statistique, en informatique et en Data Visualization. Ces compétences leur permettront de parcourir les vastes ensembles de données brutes pour en dégager les informations les plus pertinentes et les communiquer aux décideurs de leurs organisations. Les Data Scientists exploitent également l'intelligence artificielle, et plus particulièrement le Machine Learning et le Deep Learning. Ces technologies sont utilisées pour créer des modèles et réaliser des prédictions en utilisant des algorithmes et diverses techniques.

De manière générale, la **Data Science** se décompose en cinq étapes. Chacune de ces étapes requiert différentes techniques, outils et compétences.

Dans un premier temps, les données doivent être collectées, extraites à partir de différentes sources. Il s'agit ensuite de les entreposer dans une Data Warehouse, de les nettoyer, de les transformer afin qu'elles puissent être analysées. L'étape suivante est celle du traitement des données, par le biais du Data Mining (forage de données), du clustering, de la classification ou de la modélisation. Les données sont ensuite analysées à l'aide de techniques comme l'analyse prédictive, la régression ou le text mining. Enfin, la dernière étape consiste à communiquer les informations dégagées par le biais du reporting, du dashboarding ou de la Data Visualization.

5.1. Collecte des données : Recueillir les données pertinentes du domaine

À l'étape initiale de collecte des données, il est essentiel de savoir clairement pourquoi les données sont recueillies et à quoi elles serviront. Les ensembles de données pertinents résident rarement au même endroit. Certaines données peuvent provenir de systèmes de gestion de base de données relationnelle pour lesquels vous pouvez utiliser des outils comme SQL, Apache Sqoop et Java Database Connectivity pour collecter des données. Dans d'autres cas, les données peuvent provenir d'environnements Hadoop tels que Hadoop Distributed File System (HDFS), Hive, HBase et d'autres – ou de flux temps réel comme Apache Kafka, IBM Streams et les flux RSS. Quoi qu'il en soit, vous devrez prendre une décision stratégique comme le montre la figure 2 : Voulez-vous centraliser les données dans un seul endroit ou utiliser une approche fédérée ?

	Centralisée	Fédérée
Performances	Extraction de données généralement plus rapide	Délais de distribution des données plus longs
Disponibilité des données	Meilleure disponibilité	Susceptible de devenir un problème si les données ne peuvent pas être interrogées à tout moment
Coût de mise en œuvre	Combien coûte l'approche centralisée ?	Combien coûte l'approche fédérée ?
Contraintes de temps	A-t-on le temps de créer un référentiel centralisé ?	
Qualité des données	Il est plus simple de garantir la qualité des données si cette qualité est basse	

Figure 2. Comparaison des deux approches de la collecte de données : centralisée et fédérée.

Un **système centralisé** permet de résoudre les problèmes de duplication de données, d'identifier des données de base incohérentes et d'améliorer la qualité des données. Toutefois, mettre en oeuvre ce modèle peut présenter certaines difficultés, comme déterminer les emplacements géographiques, absorber le coût de l'implémentation et maintenir la conformité avec les règles métier et les réglementations.

Une **approche fédérée** vous permet d'étendre vos services pour collecter les données provenant de plusieurs sources. L'objectif est de rendre les données disponibles à tous les partenaires et unités d'une organisation. La mise en oeuvre d'un modèle fédéré se heurte à quelques obstacles, comme la synchronisation des données entre les données transactionnelles et les données de référence, les retards éventuels dans la disponibilité des données, et l'identification des rôles et des responsabilités en matière de gestion des données.

5.2. Trouver les bons outils pour chaque tâche

Quelle que soit la plateforme ou l'outil que vous utilisez, une exploration de données efficace repose sur la création d'un modèle et d'une structure de données adaptés qui peuvent être utilisés pour traiter, identifier et produire les informations dont vous avez besoin. Au moment d'examiner les différents outils de collecte de données et d'exploration de données disponibles, identifiez ceux qui peuvent être utilisés conjointement avec vos logiciels et votre infrastructure existants. Vous pouvez effectuer l'exploration de données avec des systèmes de base de données plus modestes et des outils plus simples, voire créer et écrire les vôtres ou utiliser des packs logiciels prêts à l'emploi. Vous pouvez explorer les données à partir de différents ensembles de données, y compris les bases de données SQL traditionnelles, les données textuelles brutes et les bases de données documentaires.

Données HDFS (Hadoop Distributed File System)

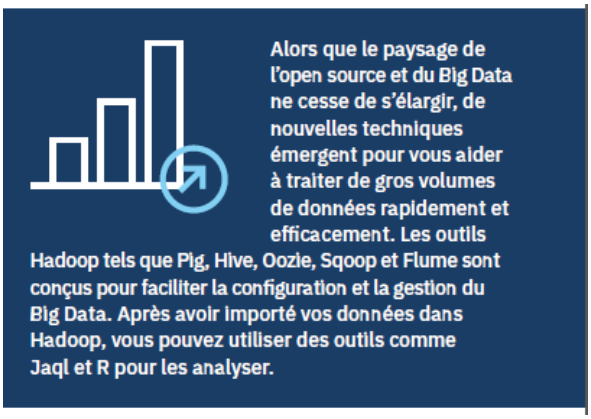
Le framework MapReduce est à la base de la plupart des processus de traitement de Big Data actuels. Créer des tâches MapReduce implique d'écrire des fonctions pour encapsuler les étapes map et reduce du calcul. Les données à traiter doivent être chargées dans le HDFS. Cette conception intégrée croît facilement, donne des options de stockage économiques et améliore les performances de votre environnement global de Big Data. Vous pouvez configurer et exécuter manuellement la migration de données, ou utiliser les outils à votre disposition. L'un de ces outils est Sqoop, qui transfère les données d'une source relationnelle vers le HDFS et vice-versa.

Flux de données

Les sources de flux de données, comme les journaux de serveur Web ou les capteurs, sont de plus en plus utilisées dans les systèmes de Big Data. Apache Kafka est devenu un outil reconnu pour le traitement de données en temps réel. Dans le cas de figure le plus simple, Kafka offre une mémoire tampon pour stocker les journaux d'application. En le combinant avec une technologie telle que Spark Streaming, vous pouvez utiliser Kafka pour suivre les modifications de données et agir vis-à-vis de ces données avant de les sauvegarder vers un emplacement final. Le mode prédictif de Kafka le rend très efficace pour détecter les fraudes, notamment lorsqu'il s'agit de contrôler la validité d'une transaction par carte de crédit en temps réel.

Apache Spark

L'un des outils open source avec la croissance la plus rapide est Apache Spark, qui présente des avantages essentiels. Tout d'abord, avec des capacités telles que le stockage en mémoire et le traitement quasi temps réel, ses performances sont nettement supérieures à celles des autres technologies de Big Data comme MapReduce ou Apache Storm. Spark prend également en charge l'évaluation paresseuse des requêtes de Big Data, ce qui aide à optimiser les différentes étapes du processus de traitement des données.



5.3. Préparation des données : Extraire et formater

Les ingénieurs des données cherchent souvent à fournir des informations exploitables en combinant données structurées et non structurées. Malheureusement, on ne peut pas présumer si ces données – même structurées – ne présentent aucun défaut et sont prêtes à l'emploi. Cela nécessite un travail préalable :

Les données peuvent être incomplètes, avec des valeurs manquantes ou incorrectes.

Les données peuvent être endommagées ou incohérentes, avec des lignes discontinues ou des champs mal placés. Les données peuvent avoir besoin d'être transformées ou

normalisées dans un format lisible par des algorithmes ou des modèles. Les données peuvent comporter trop de « bruit » : des données aléatoires et non pertinentes perturbent l'analyse.

Vous devez d'abord capturer et préserver la structure des données au moment de les transférer vers votre environnement Hadoop. Ensuite, il vous faudra transférer vos métadonnées, notamment celles relatives aux termes clés, à la sémantique et au contexte métier.

L'objectif est de mettre en relation différents ensembles de données pour en tirer des informations. Une approche possible est l'étiquetage de données. Étiqueter les données avant de les transférer vers Hadoop, Spark ou tout autre moyen d'implémentation pour analyse peut vous faire gagner beaucoup de temps. L'astuce est de traiter ces données « suffisamment » pour qu'elles soient identifiables, accessibles et couplées avec d'autres données lorsqu'elles sont chargées dans un environnement de Big Data. Si vous devez rendre plus tard vos analyses opérationnelles, il aura été judicieux de bien préparer vos données à cette étape.

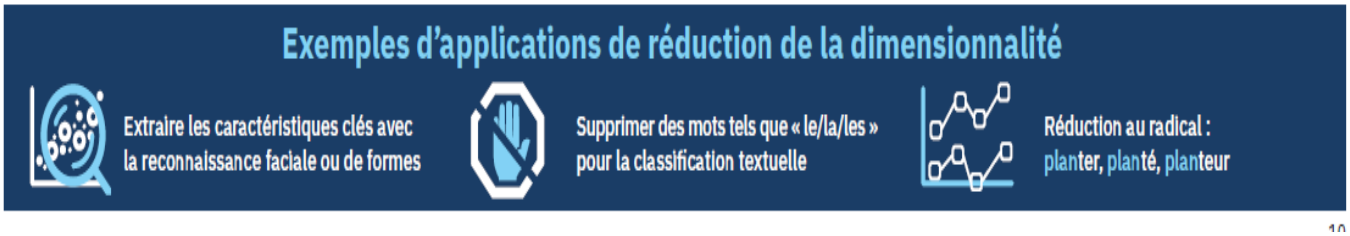
5.4. Réduire la dimensionnalité des données

La récente explosion de la taille des ensembles de données – en nombre d'enregistrements et d'attributs – a entraîné une augmentation de l'utilisation de procédures de réduction de la dimensionnalité des données. Cette réduction aide non seulement à accélérer l'exécution des algorithmes, mais aussi à améliorer les performances des modèles. Supprimer les attributs d'entrée non informatifs – ou pire encore, trompeurs – peut aider les ingénieurs à construire un modèle avec des régions de données plus étendues, des règles de classification plus générales et une meilleure performance globale sur les nouvelles données.

L'idée derrière cette réduction de la dimensionnalité est que les données brutes sont généralement constituées de deux sous-composants : des caractéristiques utiles et du bruit (structures aléatoires et non pertinentes). L'objectif est d'extraire les caractéristiques saillantes et informatives des données d'entrée afin de les utiliser dans des algorithmes prédictifs. Il existe de nombreux exemples d'ensembles de données multi-dimensionnels difficiles à traiter en une fois : chiffres manuscrits, caractéristiques faciales, journaux utilisateur, et même les expressions génétique.

Les méthodes d'extraction des caractéristiques cherchent à transformer la représentation de l'espace vectoriel du document dans une dimensionnalité moindre. Vous pourriez appliquer

une réduction au radical en vue d'extraire la racine d'un mot, et ainsi éliminer les variantes issues des différentes occurrences du même mot sous plusieurs formes grammaticales ; par exemple, « planter », « planté » et « planteur » ont tous la même racine : « plant ».



L'un des algorithmes les plus couramment utilisés pour réduire la dimensionnalité des données est l'**analyse en composantes principales (ACP).** L'objectif de l'ACP est de projeter les données d'entrée sur un sous-espace de dimension inférieure, en préservant autant de variance dans les données que possible. Les résultats d'une ACP sont généralement discutés en termes de scores de composante, parfois appelés scores factoriels (les valeurs des variables transformées correspondant à un point de données), et de chargements (le poids par lequel chaque variable originale normalisée doit être multiplié pour obtenir le score de composante).

Une autre méthode de réduction de la dimensionnalité est la **décomposition en valeurs singulières (SVD),** qui cherche à identifier et ordonner les dimensions selon les données qui présentent le plus de variations. Une fois que vous avez identifié où il en existe le plus, vous pouvez trouver la meilleure approximation des points originaux en utilisant moins de dimensions. SVD est idéal pour les applications de traitement en langage naturel car vous pouvez ignorer les variations au-dessous d'un certain seuil afin de réduire massivement vos données, tout en restant assuré que les principales relations d'intérêt ont été préservées.

Le choix de la méthode de réduction de la dimensionnalité dépend surtout des exigences de l'application qui utilisera la classification textuelle. Par exemple, certaines applications exigent une grande précision quel que soit le temps de calcul, tandis que d'autres peuvent sacrifier un certain degré de précision en faveur du temps d'exécution et du stockage requis

5.5. Transformer les données pour répondre aux exigences des algorithmes

Une exploration de données efficace exige de créer un modèle et une structure de données appropriés qui peuvent être utilisés pour traiter, identifier et générer les informations dont vous avez besoin. Quel que soit le format des données source, vous devez structurer et organiser les

informations dans un format qui prend en charge le modèle d'apprentissage automatique le plus efficace possible. Dans certains cas, il est nécessaire de transformer les données pour répondre aux exigences des algorithmes. Il existe plusieurs techniques qui permettent d'accomplir cette tâche.

La tokenisation consiste à diviser un texte en mots, phrases, symboles ou autres éléments significatifs appelés tokens. La liste de ces tokens devient une entrée d'un traitement ultérieur (ex. : analyse syntaxique ou exploration de texte). La tokenisation sémantique dépend souvent d'heuristiques simples : La ponctuation et les espaces peuvent être inclus ou non dans la liste de tokens. Toutes les chaînes contiguës de caractères alphabétiques ou numériques forment un token. Les tokens sont séparés par des espaces (ex. : saut de ligne) ou des caractères de ponctuation.

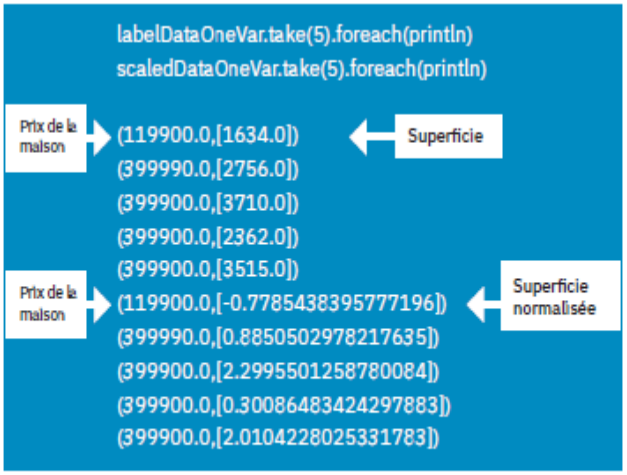


Figure 3. Exemple de standardisation des caractéristiques.

La bucketisation est un processus qui permet de définir plusieurs groupements d'enregistrements sur la base de leurs valeurs sensibles. Les valeurs sensibles apparentes des attributs sont identifiées et triées en ordre croissant selon les fréquences des attributs. Après ce tri, les valeurs sont groupées dans des buckets similaires. Seuls ceux qui contiennent des valeurs sensibles distinctes sont conservés.

La standardisation des attributs vous aide à compenser les différences dans les distributions d'attributs continus qui pourraient induire en erreur certains algorithmes de modélisation (Figure 3). La normalisation des données implique généralement de les transformer de façon à ce que chaque vecteur ait une norme unitaire. Étant donné que certains modèles s'effondrent à la valeur zéro, choisissez une plage arbitraire entre 0,1 et 0,9.

Le dummy encoding est utilisé lorsque les valeurs catégorielles ne peuvent pas être converties en valeurs numériques. Il offre un moyen d'utiliser des variables catégorielles prédictives dans plusieurs types de modèles d'estimation, comme la régression linéaire (Figure 4). Les variables muettes sont des variables « proxy » ou des équivalents numériques de faits qualitatifs dans un modèle de régression.

La normalisation des données implique généralement de les transformer de façon à ce que chaque vecteur ait une norme unitaire. Étant donné que certains modèles s'effondrent à la valeur zéro, choisissez une plage arbitraire entre 0,1 et 0,9.

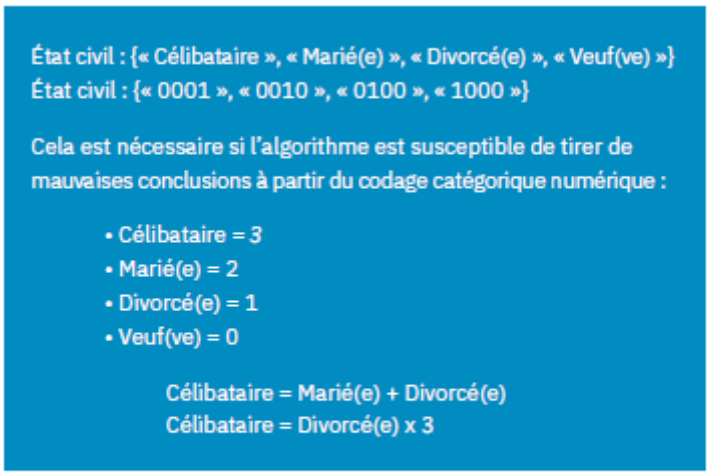


Figure 4. Les variables muettes remplacent les faits qualitatifs utilisés pour trier les données dans des catégories mutuellement exclusives (ex. : marié(e), célibataire, etc.).

Le dummy encoding est utilisé lorsque les valeurs catégorielles ne peuvent pas être converties en valeurs numériques. Il offre un moyen d'utiliser des variables catégorielles prédictives dans plusieurs types de modèles d'estimation, comme la régression linéaire (Figure 4). Les variables muettes sont des variables « proxy » ou des équivalents numériques de faits qualitatifs dans un modèle de régression.

La normalisation des données implique généralement de les transformer de façon à ce que chaque vecteur ait une norme unitaire. Étant donné que certains modèles s'effondrent à la valeur zéro, choisissez une plage arbitraire entre 0,1 et 0,9.

Le dummy encoding est utilisé lorsque les valeurs catégorielles ne peuvent pas être converties en valeurs numériques. Il offre un moyen d'utiliser des variables catégorielles prédictives dans plusieurs types de modèles d'estimation, comme la régression linéaire (Figure 4). Les variables muettes sont des variables « proxy » ou des équivalents numériques de faits qualitatifs dans un modèle de régression.

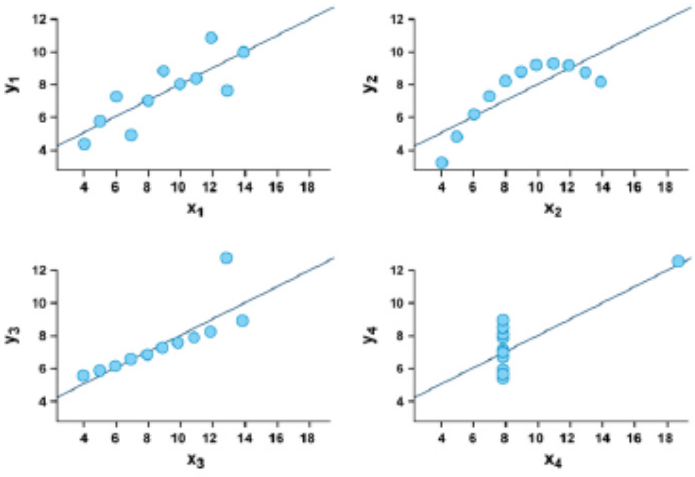
5.6. Analyse des données : Conduire les analyses avec une modélisation efficace

La phase d'analyse de données permet d'améliorer la modélisation et de la rendre plus pertinente. La modélisation met l'accent sur le développement de modèles prédictifs ou descriptifs selon l'approche analytique précédemment définie. Généralement, le processus de modélisation est extrêmement itératif, car vous commencez avec la première version de l'ensemble de données préparé, obtenez des analyses intermédiaires et utilisez-les pour affiner la préparation de données et les spécifications du modèle. La phase d'analyse de données tire profit d'outils mathématiques tels que les statistiques, les corrélations et les tests du khi-carré,

ainsi que les visualisations, qui peuvent servir à communiquer les grandes tendances au sein de l'organisation.

Une description statistique permet de décrire un ensemble de données vaste et complexe à l'aide de quelques chiffres clés. Mais il est dangereux de ne compter que sur les descriptions statistiques et d'ignorer la distribution globale. Le quartet d'Anscombe est un exemple classique de ce risque (Figure 5). Il est composé de quatre ensembles de données de 11 points x,y qui présentent des propriétés statistiques simples quasiment identiques, mais qui apparaissent très différents lorsqu'ils sont représentés sous forme graphique.

Les descriptions statistiques pouvant prêter à confusion, il est important de les utiliser comme simple outil dans un processus plus large d'analyse de données. Visualiser les données vous permet de réexaminer les descriptions statistiques et de les recontextualiser si nécessaire.



Les quatre ensembles de données ont des propriétés statistiques :

- La moyenne de x est 9
- La variance de x est 11
- La moyenne de y est env. 7,50
- La variance de y est env. 4,12
- La corrélation est 0,816

La droite de régression linéaire est env. y = 3,00 + 0,500 x

Figure 5. Exemple de quartet d'Anscombe.

Rendre accessibles les données complexes

Un des principaux objectifs de la visualisation des données est de communiquer l'information de façon claire et efficace au moyen de graphiques statistiques, de tracés et d'infographies. Les données numériques peuvent être représentées à l'aide de points, de lignes ou de barres pour communiquer visuellement un message quantitatif.

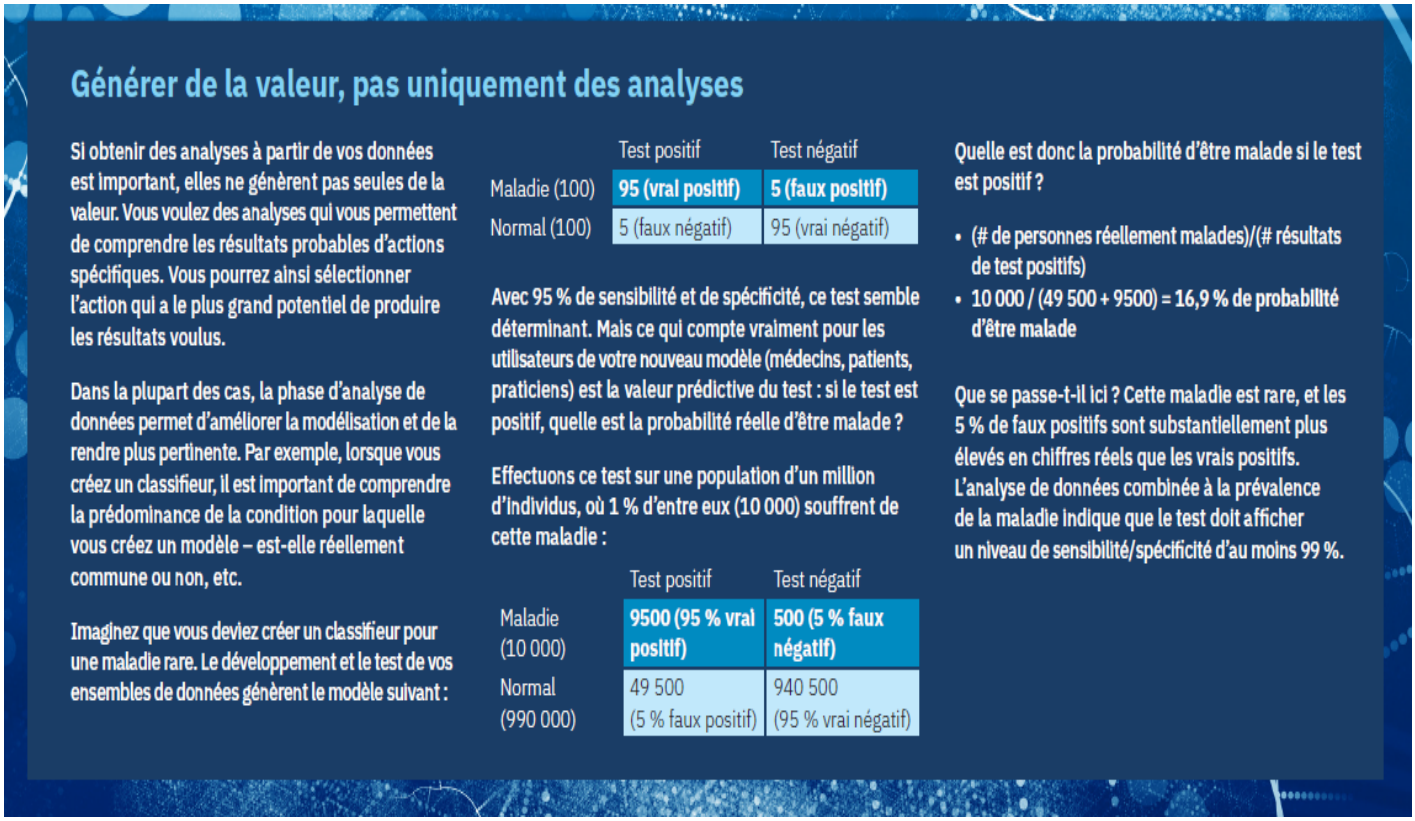
Une visualisation efficace aide les utilisateurs à analyser et à raisonner sur les données et les informations. Elle rend des données complexes plus accessibles, compréhensibles et exploitables. Le principe de conception du graphique (comment montrer les comparaisons ou les causalités) devrait suivre la tâche d'analyse. La cartographie en particulier est un outil extrêmement utile pour rendre les données accessibles. Si vous pouvez visualiser diverses

données grâce à la cartographie, vous pouvez souvent faire votre analyse en deux fois moins de temps. Même des utilisateurs qui ne possèdent pas de grandes compétences en mathématiques peuvent suivre ce processus et atteindre leurs objectifs plus rapidement.

L'une des méthodes de visualisation les plus efficaces – surtout avec des données géographiques du type latitude ou longitude – consiste à afficher un échantillon des données sur une carte. Cela permet de voir les tendances ou parfois des erreurs évidentes, comme des valeurs de latitude et de longitude avec le mauvais signe (points de données dans l'hémisphère sud au lieu de l'hémisphère nord et vice-versa). Les lignes et les barres sont également utiles pour identifier visuellement des tendances telles que des corrélations positives ou négatives.

Une visualisation efficace aide les utilisateurs à analyser et faire sens de données et d'informations collectées. Elle rend des données complexes plus accessibles, compréhensibles et exploitables.

Pour une technique donnée, vous pouvez essayer plusieurs algorithmes et leurs paramètres respectifs afin de trouver le meilleur modèle pour les variables disponibles. Vous pouvez aussi affecter des tests de signification statistique au modèle comme preuve supplémentaire de sa qualité. Celle-ci peut être utile pour justifier la mise en oeuvre d'un modèle ou pour prendre des décisions lorsque les enjeux sont élevés, comme un protocole médical supplémentaire coûteux ou un système de vol critique.



6. Les cas d'usage et applications

Les cas d'usage de la Data Science sont aussi nombreux que variés. Cette technologie est utilisée pour **assister la prise de décision en entreprise**, mais permet aussi l'automatisation de certaines tâches.

Elle est utilisée à des fins de **détection d'anomalies ou de fraude**. La science des données permet aussi la classification, par exemple pour trier automatiquement les emails dans votre boîte.

Elle **permet aussi la prédiction**, par exemple pour les ventes ou les revenus. En l'utilisant, il est possible de détecter des tendances ou des " patterns " dans les ensembles de données.

La Data Science se cache aussi derrière les technologies de **reconnaissance faciale, vocale ou textuelle**. Elle alimente aussi les moteurs de recommandations capables de vous suggérer des produits ou du contenu en fonction de vos préférences.

D'un secteur d'activité à l'autre, la Data Science est exploitée de différentes manières. Dans **le domaine de la santé**, les données permettent aujourd'hui de mieux comprendre les maladies, de recourir à la médecine préventive, d'inventer de nouveaux traitements ou d'accélérer les diagnostics.

En **logistique**, la Data Science aide à optimiser les itinéraires et les opérations internes en temps réel en tenant compte de facteurs comme la météo ou le trafic. Dans **la finance**, elle permet d'automatiser le traitement des données d'accords de crédit grâce au Traitement Naturel du Langage (Vous n'êtes pas familier avec ce concept, découvrez le NLP dans notre article dédié) ou de détecter la fraude grâce au Machine Learning.

Les **entreprises de retail** l'utilisent pour le ciblage publicitaire et le marketing personnalisé. Les moteurs de recommandations, basés sur l'analyse des préférences du consommateur, sont utilisés par Google pour son moteur de recherche web, par les plateformes de streaming comme Netflix ou Spotify, et par les entreprises de e-commerce comme Amazon.

Les **entreprises de cybersécurité** se tournent vers l'IA et la science des données pour découvrir de nouveaux malwares au quotidien. Même les voitures autonomes reposent sur la Data Science et l'analyse prédictive pour ajuster leur vitesse, éviter les obstacles et les changements de voie dangereux ou choisir l'itinéraire le plus rapid