Entreposage et fouille de données

Synthèse

Maya Besma Le Corre [mai 2018]

«Tous les modèles sont faux, certains sont utiles» Georges Box

Nadib Bandi – Sculpture abstraite - Installation graffiti en 3D

Contenu

Préambule

Ce document est une synthèse personnelle du cours dispensé par le CNAM et intitulé "UESTAT211 Entreposage et fouille de données". Cet enseignement a été suivi dans le cadre de la certification "Analyste de données massives" par la promotion d'entreprise ProBTP de oct. 2017 à mars 2018. Les principales méthodes de fouille de données sont résumées ici dans une liste non exhaustive. L'objectif de ce document est de fournir une vue d'ensemble d'un projet de datascience en entreprise.

Introduction

La gigantesque masse de données hétérogènes, volatiles et d'origines multiples représente le continent des explorateurs que sont les datascientists. Le forage de montagnes de données à l'aide de traitements automatisés et massivement parallèles mobilisent le monde numérique à l'image de la ruée vers l'or Californienne. Même si les techniques d'apprentissage semblent nouvelles, les experts savent que la statistique inférentielle date. La pratique inductive depuis un échantillon de la population, consciencieusement élaboré, n'est pas révolue. Au contraire, le volume à traiter favorise son utilisation dans les techniques de passage à l'échelle.

En effet, les méthodes statistiques et les bases mathématiques qui soutiennent les techniques d'analyse de données massives n'ont pas connues de révolution scientifique autre que la puissance accrue des calculateurs. En cumulant les avancées technologiques et l'avènement d'un volume massif de données presque gratuites, les mineurs des temps modernes adoptent une démarche analytique rigoureuse et pragmatique, pour extraire de la masse, des connaissances bénéfiques et exploitables pour l'entreprise. La recherche en intelligence artificiel plus récente a permis de produire des algorithmes puissants de dataming sur les données multimédias en particulier. Désormais, il est plus aisé d'envisager l'intégration d'un processus de datamining en production.

Cette synthèse s'attache à décrire le déroulement d'un projet de datascience en passant en revue le panorama des principales méthodes statistiques dans l'idée de fournir un aperçu de la discipline. Cette approche pratique implique une structuration du plan en étape de projet et amène à regrouper les différentes méthodes en fonction d'objectifs d'analyse. La démarche générale de fouille de données consiste à définir l'objectif du projet et à comprendre les données pendant la phase exploratoire. S'en suit une phase de modélisation incluant la validation du modèle. Certains modèles ont vocation à poursuivre leur chemin vers les systèmes de production avec la phase de déploiement.

1 Phase exploratoire

Dans cette phase on veut comprendre le besoin métier et les données. En général, il faut reformuler la question posée par le "client" en un objectif d'analyse de données et choisir des méthodes analytiques. Ces choix dépendent autant de l'objectif de l'étude que des données disponibles. Faut-il décrire ? Prédire ? Expliquer ? Ou une combinaison de ces questions ? Une fois la problématique identifiée il faut aller explorer les données et donc les décrire ensuite si le besoin est de prédire alors on va bâtir un modèle et le déployer.

1.1 Processus de fouille de données

Un projet de fouille de données est un processus en plusieurs étapes. Certaines d'entre elles représentent une charge de travail qui peut être conséquente et d'autres qui sont beaucoup moins chronophage mais beaucoup plus décisives. Les étapes importantes sont la définition de l'objectif et la validation du modèle. L'étape la plus chronophage est la préparation des données. Le schéma suivant synthétise les principales étapes d'un processus de fouille de données.

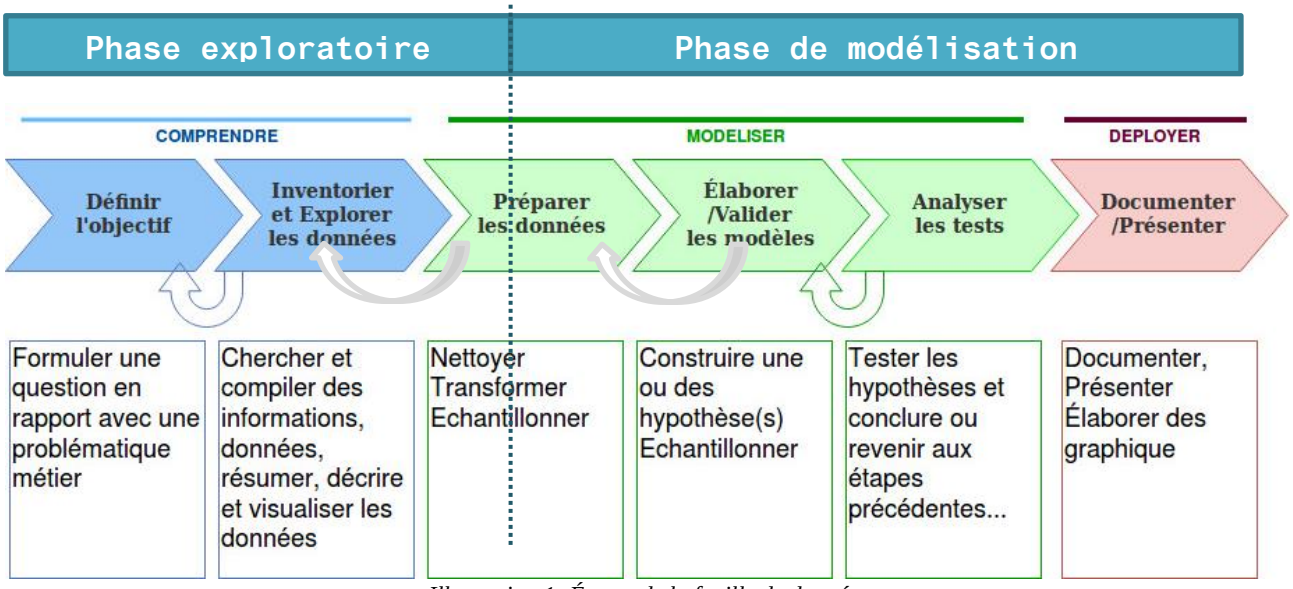


Illustration 1: Étapes de la fouille de données

1.2 Définir l'objectif

Le but est de découvrir de la connaissance au-delà des évidences. Il convient alors de reformuler le problème posé de manière à pouvoir y répondre à l'aide de méthodes de fouille de données. Cette étape est cruciale et contribue de manière évidente au succès du projet.

1.2.1 Quelques familles d'objectif

Le panorama de méthodes listées dans ce document doit pouvoir répondre à ces quelques objectifs formulés ici de manière générique. Il se peut qu'à la fin de la phase exploratoire l'objectif soit révisé.

Décrire	Résumer les données de la manière la plus intelligible en gardant le maximum d'information.
Prédire	> Expliquer la relation d'une variable avec les autres variables et prédire la valeur. > Prédire la valeur sans expliquer la relation entre les variables. > Une fois le modèle établi, on peut aussi vouloir fiabiliser la prédiction par l'apprentissage statistique (en utilisant les méthodes d'apprentissage pour estimer au mieux les paramètres du modèle). Les objectifs dans le domaine de la prédiction peuvent être de natures différentes : soit une classe d'appartenance (au moins 2 classes) soit une quantité
Structurer	> Décrire la structure des données : Trouver ou établir des groupes de telle sorte que les individus au sein d'un même groupe se ressemblent le plus possible et que les individus d'un groupe à l'autre soient les plus différents possible. On cherche à savoir comment les données sont organisées.
Associer	> Trouver les ensembles de valeurs de variables qui soient le plus corrélées > Chercher les similarités pour un individu donnée avec les autres individus par rapport aux contenus des données
Rechercher les outliers	Détecter les anomalies, les comportements inhabituels, les événements rares, ...

1.2.2 Choix de la méthode

Le choix des méthodes dépend de la nature des données et surtout du besoin. Dans la phase exploratoire l'objectif est de découvrir alors que dans la phase de modélisation l'objectif est de fournir un modèle. L'exploration comme la modélisation s'appuient sur des méthodes.
On peut les regrouper en quatre grandes familles (toutes sortes de découpage sont proposés par différents auteurs, celle-ci n'engage que son auteure) :

- les méthodes d'analyse descriptive visant à résumer l'information, à réduire le nombre de dimension,
- les méthodes d'analyse de structure visant à résumer l'information de façon logique ou graphique,
- les méthodes explicatives visant à prédire une valeur cible avec un résumé de l'information lisible,
- les méthodes prédictives visant à prédire sans information explicite sur le résumé.

En phase exploratoire, les méthodes d'analyse descriptives sont choisies en fonction des types de données et sont principalement des méthodes d'analyse factorielle mais on pourrait également si le but est de décrire des groupes utiliser une méthode d'analyse de structure (clustering ou réduction de dimension).

En phase de modélisation, on choisit des méthodes en fonction de l'objectif :
- Soit il s'agit d'établir un regroupement d'individus on sélectionne des méthodes d'analyse de structure,
- Soit on veut pouvoir expliquer la relation entre les variables, on emploi dans ce cas des méthodes explicatives pour analyser la relation de dépendance, entre la variable à expliquer et les variables explicatives,
- Soit on veut pouvoir prédire sans nécessairement expliquer, alors on choisit parmi les méthodes prédictives.

Par soucis d'organisation dans la galerie de méthodes, sont listées en phase exploratoire des méthodes d'analyse factorielle essentiellement et en phase de modélisation les méthodes visant à produire un modèle y compris les modèles descriptifs. Un diagramme résumant des critères «objectifs» de choix de méthode figurent en annexe du document (c'est-à-dire non lié aux uses et coutumes de chaque secteur).

1.3 Comprendre les données

C'est la phase la plus chronophage dans un projet avec de nombreux pièges à éviter qui peuvent amener le projet dans la mauvaise direction. L'utilisation de techniques statistiques assistée par des outils "intelligents" qui prémâchent le travail présente un risque pour les débutants, celui de ne pas utiliser les techniques appropriées ou de ne pas maîtriser les transformations de données effectuées. La compréhension des données favorise l'obtention de résultats interprétables.

1.3.1 Préparer les données

Les données collectées doivent être nettoyées, transformées, simplifiées. Le prétraitement des données va chercher à améliorer la qualité des données. Il faut mettre en œuvre un ensemble de techniques d'appréhension des valeurs manquantes ou aberrantes et des techniques de transformation (agrégations, etc.)

1.3.1.1 Collecte des données

- Données internes à l'entreprise :

Contrairement à la statistique classique, la collecte de données n'est pas la préoccupation majeure dans la mesure où l'exercice consiste à exploiter une masse de données déjà stockée. La problématique est liée au fait que la donnée est structurée de manière optimisée pour le stockage mais pas pour un usage analytique. Les difficultés peuvent être nombreuses : L'information est éclatée, peu documentée, tributaire de l'interprétation d'experts etc. Le travaille à cette étape consiste à évaluer si les données sont exploitables et exhaustives en vue du problème à résoudre. L'échantillonnage est par conséquent un bon moyen de maîtriser les données. Des techniques d'échantillonnage peuvent être utilisées pour éviter les problèmes de sous ou sur représentation de certains groupes d'individus.

- Open data :

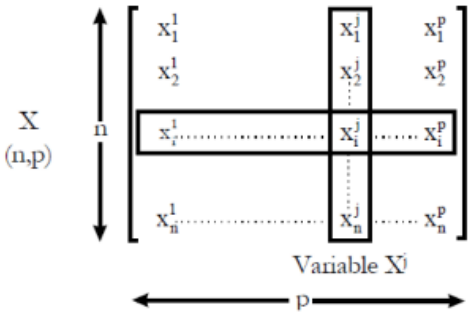
"L'open data ou donnée ouverte est une donnée numérique dont l'accès et l'usage sont laissés libres aux usagers. Elle peut être d'origine publique ou privée, produite notamment par une collectivité, un service public (éventuellement délégué) ou une entreprise. Elle est diffusée de manière structurée selon une méthode et une licence ouverte garantissant son libre accès et sa réutilisation par tous, sans restriction technique, juridique ou financière." (Source : Wikipédia)

C'est la possibilité de croiser des données de provenance diverses au travers de divers composants interopérables. Dans les faits, la granularité n'est pas forcement adéquate. La phase de préparation des données reste indispensable et est plutôt chronophage pour la validation des données. La documentation manque souvent de précision. La philosophie du mouvement n'est pas toujours respectée. En somme, la donnée est libre et facile d'accès mais elle est présente en trop grand nombre et le plus souvent de manière peu pertinente. L'open data n'est pas la panacée même si parfois cela est utile. Donc il faut prévoir dans le projet d'étude suffisamment de temps pour trouver la bonne source de données, voir peut être même, si le budget le permet, acheter un fichier déjà qualifié avec l'assistance d'un expert pour répondre à nos questions.

- Big Data :

Les données sont fragmentées, hétérogènes, dispersées dans différents systèmes, soumises à des réglementations pour certaines. Toutes ses caractéristiques sont autant de difficultés à surmonter dans la phase de collecte des données. Il faut envisager des scenarii de collecte multi-formats et multi-sources et imaginer comment corréler les données entre-elles pour les besoins des utilisateurs. Les projets 'Big data' en entreprise nécessitent de nombreux traitements et l'addition peut s'avérer élevée. Un calcul de retour sur investissement s'impose avant d'engager des sommes importantes. Tout comme une étude de faisabilité s'impose avant de passer trop de temps. Ces études préalables peuvent être élaborées sur un échantillon dans un premier temps pour évaluer le potentiel commercial ou administratif d'un modèle. Un projet Big data reste un projet et cette typologie de projet ne modifie pas les règles de bon sens.

1.3.1.2 Les différents jeux de données



Une analyse rapide permet de repérer les cas de figures récurrents ou atypiques : nombre de lignes et de colonnes, taux de remplissage des variables, valeurs extrêmes, et nombre de tableaux. On peut se retrouver face à un jeu de données qui contient peu d'individus et beaucoup de variables ou à l'inverse peu de variables et beaucoup d'individus. On peut aussi avoir beaucoup de données manquantes. Ces jeux de données présentent des caractéristiques particulières qu'il faut considérer dans le choix des méthodes.

- **Le jeu de données de base**

La structure de base utilisée pour la fouille de données est un tableau à deux dimensions avec les individus (objets, observations, etc.) en ligne et les variables (caractères, attributs, etc.) en colonne. Les valeurs du tableau sont des données nominales, ordinales, discrètes ou continues. De manière générale, on note N le nombre d'individus (nombre de lignes) et P le nombre de variables (nombre de colonnes). Le cas général admet l'hypothèse suivante : N > = P.

- **Tableau de contingence**

On peut avoir des lignes qui ne participent pas à la construction de l'analyse : des moyennes ou regroupement d'individus en lignes ou des variables qui sont redondantes ou très lié à d'autres variables par un calcul.

- **Un jeu de données clairsemé (sparse data en anglais)**

Quand le nombre de variables P est très supérieur au nombre d'individus N avec en plus une forte probabilité de contenir un grand nombre de valeurs nulles, on parle de matrice creuse. On considère une matrice creuse comme le représentant d'un système peu couplé. C'est le cas que l'on rencontre avec des données décrivant des réseaux par exemple où des données connectées se caractérisent par une faible densité de connexions. En régression linéaire, l'estimateur n'existe pas et la sélection de variable devient inefficace ou ingérable. Dans ce cas, on cherche à résumer l'ensemble des variables par combinaison linéaire. On va effectuer une régression sur composante principale.

- **Données évolutives**

On peut avoir des tableaux multiples à analyser. On considère chaque tableau comme un bloc. On peut analyser la structure inter-bloc grâce à la méthode d'analyse multi-bloc. Cette généralisation des méthodes applicables à un seul tableau permet d'analyser les relations inter-structures et de comprendre les relations entre les blocs. Si chaque bloc représente une évolution dans le temps, on peut étudier l'évolution temporelle des mêmes individus et des mêmes variables. On peut aussi analyser des blocs ayant des individus en commun mais des variables différentes ou des variables communes avec des groupes individus différents. Dans ce cas on étudiera les trajectoires de l'individu moyen ou celles des variables dans chaque tableau.

- **"Nouveaux" types de données**

Des méthodes spécifiques bien adaptées à certains cas de figure ne sont pas abordées ici :
- Documents textuelles
- Données multi-vues
- Images et vidéos
- Réseaux sociaux

1.3.1.3 La qualité des données

Cette étape repose sur le principe qui s'énonce ainsi "Garbage In Garbage Out". Cette expression signifie que la qualité des résultats dépend de la qualité des données en entrée. La préparation des données est donc primordiale. Il y a cinq critères principaux à observer.
- Tout d'abord il doit y avoir un rapport avec la problématique de départ (même si cela parait évident en théorie, en pratique si on est étranger au domaine métier, l'erreur est probable).
- Ensuite, le tableau de données doit avoir du sens. C'est à dire que l'ensemble des variables du tableau doit former une unité fonctionnelle cohérente. On ne peut pas tirer d'enseignement de données qui n'ont rien à voir les unes avec les autres (par exemple : la taille des individus et le numéro de la rue où ils habitent). On pourra peut - être mettre en évidence une corrélation mais l'interprétation sera bien délicate.
- Par la suite, il faut veiller à avoir un maximum de valeurs parmi les données explicatives par rapport à la variable à expliquer. Si des valeurs pour un groupe de variables viennent à manquer par rapport à un autre groupe de variables alors il devient impossible d'inférer ou de résumer l'ensemble du tableau de données. On se retrouve avec des sous tableaux que l'on peut analyser séparément mais pas dans un seul tableau.
- De même, les valeurs aberrantes doivent être étudiées pour distinguer si la variable est acceptable ou pas, ou bien si les individus peuvent être conservés ou éliminés. Cela dépendra de la problématique et des données indispensables à l'étude.
- Enfin, la taille de l'échantillon doit être suffisamment représentative de la population. En effet, on doit pouvoir généraliser le modèle. Si l'échantillon est trop petit alors la généralisation devient périlleuse.

1.3.1.4 Données manquantes

Les données manquantes impliquent forcément un manque de précision dans les résultats obtenus mais peuvent aussi introduire un biais dans l'analyse. On imputera les données manquantes de sorte qu'elles ne contribuent pas à la construction des dimensions. Lorsqu'on ne dispose pas de certaines données, on va chercher à combler les manques à l'aide de méthodes simples si l'expert métier peut donner des règles ou par méthodes statistiques (moyenne, médiane, régression locale, les K plus proches voisins, etc.). On dit qu'on impute les données manquantes.
Pour opérer de la bonne façon il convient d'en connaître la cause :
- Si le fait qu'elle soit manquante a une signification alors on peut imputer une valeur
- Si la donnée n'est pas facile à avoir ou coûte trop cher, souvent l'expert peut déterminer une règle d'imputation mais selon la sensibilité de l'information non communiquée une perte de précision pourrait être déplorée
- Si elle est manquante de façon aléatoire alors on peut ignorer les observations avec données manquantes sans biaiser le résultat
- Si l'absence n'est pas totalement aléatoire, alors des techniques permettent d'imputer de la façon la plus neutre possible. Un article complet sur les techniques se trouve ici : http://wikistat.fr/pdf/st-m-app-idm.pdf

Il faut vérifier si la méthode tolère les données manquantes avant de prendre le soin de les traiter, surtout s'il y a un risque d'introduire un biais par l'imputation.

1.3.1.5 Valeurs aberrantes

On détecte rapidement les valeurs très éloignées des autres observations dans la phase exploratoire. Si possible, en connaître la cause peut aider à décider de conserver ou éliminer ces valeurs de l'échantillon. Si on a beaucoup de valeurs aberrantes on peut remettre en cause la validité des données et remonter à la source pour s'assurer de la fiabilité des données. Il faut appliquer des méthodes pour détecter s'il s'agit réellement d'une valeur aberrante.
Les valeurs extrêmes vont avoir une importance dans la phase de création et de validation du modèle. Les calculs de tests paramétriques y sont très sensibles. Il convient donc de les traiter pour ne pas perturber la validation du modèle. Si la cause est inconnue et que le choix d'éliminer les valeurs s'avère problématique, il conviendra d'utiliser des tests non paramétriques qui sont moins sensibles à ces valeurs aberrantes. Certains modèles les tolèrent ou permettent de les mettre en évidence. Si l'objectif de l'étude porte sur ces valeurs il faut les conserver et adapter le choix des méthodes en conséquence.

1.3.1.6 Données redondantes et bruit

Quand les variables fournissent plusieurs fois la même information cela peut avoir un impact sur la modélisation. Certaines méthodes réagissent mal face à cette configuration de données. Certaines variables n'apportent aucune information. On parle de bruit. La sélection de variable est issue d'analyse de ces phénomènes soit avant l'utilisation de méthode soit par la méthode elle-même.

1.3.1.7 Transformations

Les données brutes sont des valeurs numériques ou textuelles. Les transformations peuvent porter sur le type de données à convertir, sur le nom des variables, sur les variables elles-mêmes pour traduire des modalités nominales en tableau disjonctifs complet la plupart du temps. Cela va dépendre de la méthode que l'on veut employer. Il y a quelques précautions à prendre lors de cette transformation pour ne pas corrompre les données.
Comme il est difficile de généraliser, on peut donner quelques exemples :
- modifier l'unité de mesure d'une variable pour rétablir un ordre de grandeur comparable aux autres
- pondérer les observations pour rééquilibrer les effectifs au sein des différents groupes afin qu'il soit représentatif de la population globale; On dit que l'on redresse les données
- générer une variable à partir des autres variables

- recoder une variable textuelle en code unifié
- regrouper des modalités
- codage en tableau disjonctif complet
- discrétiser une variable ordinale en classe
- ajouter un rang (variable numérique entière)
- générer des données aléatoires suivant une loi de distribution
- construire sa base de travail en agrégeant les lignes pour aboutir à une définition des individus à étudier pour une problématique donnée
- éliminer les variables et les lignes redondantes
- appliquer une fonction de transformation (exemple une fonction logarithme ou une fonction exponentielle)

1.3.2 Démarche exploratoire

Une fois les données prêtes pour l'analyse exploratoire on procède par étape de découverte. On va mobiliser toutes les techniques de la statistique descriptive et principalement les méthodes d'analyse factorielle. En premier lieu on calcul des paramètres statistiques pour visualiser la tendance centrale, la position et la dispersion des valeurs pour chaque variable prise indépendamment. C'est l'étape d'analyse uni-variée. Ensuite, on étudie deux à deux les variables pour caractériser l'existence d'un lien en évaluant la corrélation entre des valeurs quantitatives et l'indépendance entre des valeurs qualitatives. Enfin, on étudie le tableau dans son ensemble en le transformant pour l'étudier plus facilement dans un espace de plus faible dimension.

1.3.2.1 Analyse uni-variée

L'analyse uni-variée cherche à résumer la répartition et la densité de chaque variable du tableau. Un jeu de donnée est pertinent s'il existe de la variance, dans le cas contraire on ne peut rien déduire tant en classification qu'en régression. Donc une variable qui contient une seule valeur peut être éliminée d'emblée car elle n'apporte pas d'information supplémentaire. On observe la dispersion avec l'amplitude des valeurs, la variance et l'écart-type qui mesure la dispersion autour de la moyenne. Quand la moyenne est égale à la médiane, la série est parfaitement symétrique.
On va résumer les données en calculant des valeurs sur les variables quantitatives : moyenne, écart-type, quantiles à minima. On représente ces valeurs à l'aide de boites à moustache. C'est une représentation graphique de la répartition des valeurs d'une variable quantitative. Elles sont particulièrement utiles pour comparer les distributions de plusieurs variables ou d'une même variable entre différents groupes. On peut poursuivre l'analyse en calculant le coefficient de dispersion (écart-type divisée par la moyenne) sur chaque variable. Si on veut mesurer l'inégalité de la dispersion on peut calculer l'indice de GINI. Cet indice fournit une information sur la concentration d'une valeur dans une série. Ce coefficient permet de comparer la dispersion de plusieurs variables entre elles si elles n'ont pas la même unité de mesure.

En ce qui concerne les variables qualitatives, on pourra s'attacher à calculer pour chaque modalité l'effectif et la fréquence. Si le nombre de modalité est trop important on groupera les modalités peu fréquentes sous une nouvelle valeur. On utilise la représentation en camembert ou diagramme en bâton pour les effectifs.

Les logiciels fournissent toute la panoplie de calculs statistiques (moyenne, quartiles, effectif, fréquence, etc.) qui doivent être utilisés selon le besoin. Le choix de la représentation graphique doit être adapté à l'étude menée. On peut choisir de représenter l'effectif seul ou l'effectif cumulé. On ne montrera pas la même chose. Cette étape peut amener à sélectionner les variables à conserver pour l'étude mais certaines méthodes le font d'elle-même alors si on le fait manuellement il faut être certain que les variables non prise en comptes n'ont réellement aucun intérêt et que nous les ignorons pas à cause d'un mauvais à priori. De manière générale, on préfère garder le plus possible de variables.

1.3.2.2 Analyse bi-variée

On analyse en particulier les dépendances des variables conjointement deux à deux. De manière générale, on cherchera à centrer et réduire les variables pour comparer plus facilement les variations sans unité et sans effet d'échelle. Pour aller vite, on représente une matrice de nuage de points avec les paires de variables. On établit également un diagramme de la distribution des variables. On analyse la distribution en cherchant à se rapprocher d'une loi de probabilité. On peut déduire si la relation entre les variables traduit une fonction paramétrique ou non. Pour une variable 'normale' on considère que l'intervalle défini par l'écart-type comprend la plupart des observations. On recherche les corrélations entre variables quantitatives. La mesure de la dépendance linéaire est fournie par le coefficient de corrélation. Il est égal à 1 dans le cas où l'une des variables est une fonction affine croissante de l'autre variable, à -1 dans le cas où une variable est une fonction affine et décroissante. Les valeurs intermédiaires renseignent sur le degré de dépendance linéaire entre les deux variables. On peut faire apparaître la force des dépendances conjointes au sein d'une matrice de corrélation. Le diagramme correspondant est le corrélogramme. Ceci va permettre de montrer l'existence une relation linéaire et donc permettre de choisir des tests paramétriques.
Il est important de rappeler pour l'interprétation que la corrélation n'implique pas la cause (*"Cum hoc ergo propter hoc"* en latin). Pour les variables qualitatives, on cherche la relation de dépendance en effectuant le test d'indépendance (Khi²). Pour cela on doit élaborer des tableaux croisés. S'il y a indépendance, on estime que l'on ne peut pas interpréter. En revanche, quand l'hypothèse d'indépendance peut être rejetée on pourra visualiser la représentation de l'indépendance dans un stéréogramme. Les graphiques en mosaïque ou des diagrammes en barre cumulées sont des représentations adaptées pour les restitutions de ces résultats. Si on veut comparer deux variables ayant chacune deux modalités alors le test de Fisher est approprié. Les boites à moustache peuvent aussi nous aider à comparer des valeurs quantitatives et qualitatives dans une analyse bi-variée. On étudie la relation entre une variable qualitative et une variable quantitative avec la méthode ANOVA. On analyse la covariance avec la méthode ANACOVA.

1.3.2.3 Analyse multi-variée

Le choix de la méthode (analyse factorielle ou de clustering) dépend des types de variables contenues dans le tableau de données (quantitatives, qualitatives ou mixtes) mais surtout de l'objectif. Si l'objectif est d'obtenir une vision globale du jeu de données en cherchant des corrélations ou des dépendances entre variables (pas de variable à expliquer) alors on doit employer les méthodes factorielles.

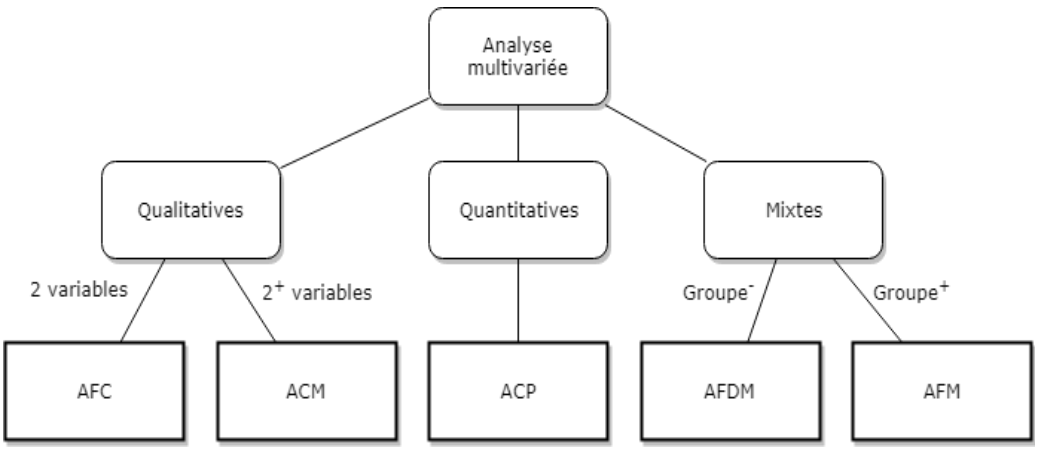


Illustration 2: Choix de la méthode d'analyse factorielle

On effectue une ACP quand le tableau contient des variables quantitatives, une ACM pour des variables qualitatives (AFC pour deux variables catégorielles ou sur un tableau de contingence) et une AFDM si le tableau contient des données mixtes (AFM si les données sont structurées en groupe).

Si l'objectif est de trouver la structure sous-jacentes des données alors les techniques de clustering permettent de déceler les proximités entre individus ou entres variables.
Il existe également une méthode pour les besoins de classification (classement des individus défini par des variables qualitatives dans des catégories). Il s'agit de l'AFD.

- **L'approche factorielle**

L'analyse factorielle est une démarche qui transforme des variables liées entre elles en nouvelles variables décorrélées les unes des autres. Ces nouvelles variables sont nommées dimensions et sont représentées comme des axes dans l'espace à plusieurs dimensions. Ceci permet de réduire le nombre de variables à étudier et de supprimer la redondance d'information qui peut nuire à l'interprétation. La transformation permet d'étudier la forme du nuage de points en réalisant une projection de ces points sur un plan. On retient la (ou les) photo(s) qui représente(nt) le mieux le nuage de point. On va hiérarchiser l'information en ordonnant les axes en commençant par celui qui contient le plus de variance vers celui qui en contient le moins. Cette transformation s'effectue à l'aide d'opérations algébriques sur la matrice de donnée appelée X.

- **Les étapes de transformations algébriques en quelques lignes**

1. centrer et réduire X
2. calculer la matrice d'inertie associée
3. diagonaliser cette dernière
4. effectuer le changement de base avec les valeurs propres obtenues
5. projeter les points sur les axes d'inertie du nuage de point

En sortie, on obtient les coordonnées des variables, une mesure des angles (Cos^2), une mesure de leurs contributions aux dimensions.
Il y a des notions essentielles à considérer dans le traitement :

- La notion de distance dont le mode de calcul influence fortement les résultats de l'analyse et l'inertie du nuage de point (la dispersion autour du centre de gravité)
- La réduction unifie l'importance des variables quelques soit leurs dispersions et neutralise les différentes unités de mesures, ce qui permet d'interpréter et de :
  - visualiser les individus avec une notion de distance
  - visualiser les corrélations entre variables/modalités

- **Interprétation**

- Description des dimensions et de l'inertie : C'est la même méthodologie dans toutes les techniques d'analyse factorielle. L'interprétation du résultat consiste à retenir un nombre de dimensions pour lesquelles on mesure la qualité des représentations. Il y a plusieurs façons de retenir les p premiers axes (en général maximum 3), la plus visuelle est la technique du "coude" (inflexion brusque de la courbe). On décrit ces axes factoriels dans l'ordre des inerties croissantes. Il faut décrire chaque axe à partir des variables qui contribuent le plus dans sa construction en observant le diagramme des valeurs propres. La compression de l'information se mesure par la somme des pourcentages d'information expliquée par chaque axe. On mesure ainsi la quantité d'information représentée (diagramme : scree plot).

- Qualité de représentation : On mesure la qualité de projection sur le plan à partir de l'angle formé entre l'élément et le plan de projection. Seuls les éléments correctement projetés sont interprétables. Dans la représentation du cercle de corrélation, les variables les plus proches du cercle sont les mieux représentées.

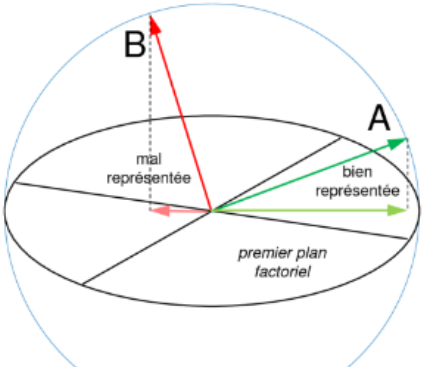


Illustration 3 : Contribution relative d'un axe dans l'explication de l'inertie d'une variable

■ **Les différentes techniques**

- ACP (Analyse en Composantes Principales)
  - Input : Tableau de variables quantitatives
  - Output : Nuage des individus et cercle des corrélations entre variables

  - Étapes d'analyse :
    1. Choisir les variables actives
    2. Choisir de réduire ou non les variables (obligatoire si les unités des variables sont différentes sinon selon le choix de modèle voulu par rapport à la variance des variables et à la notion de bruit)
    3. Réaliser l'ACP
    4. Choisir le nombre de dimension à interpréter
    5. Interpréter simultanément le graphe des individus et les graphes des variables
    6. Utiliser des indicateurs pour enrichir l'interprétation (vérification mathématiques des proximités visibles sur le plan)
    7. Revenir aux données brutes ou aux données centrées réduites pour interpréter (on valide notre interprétation)
    8. On peut ajouter des variables supplémentaires (variables retirées dans la 1ere ACP) et refaire une ACP il s'agit de l'analyse externe.

  - Interprétation : Pour le cercle de corrélation il faut donner un sens à chaque axe factoriel à partir des corrélations et des oppositions entre variables. Pour le nuage des individus, les proximités des individus projetés sur les plans traduisent des ressemblances par rapport aux variables.

- ACM (Analyse des Correspondances Multiples)
  - Input : Tableau de variables qualitatives (exemple classique : résultat d'enquête)
  - Output : représentation simultanées des individus et des modalités sur le plan et une structure hiérarchisée des données

  - Traitement : Recoder en données quantitatives les modalités de chaque variable qualitative. Tableau disjonctif complet (TDC) ou tableau de burt qualitatives ou tableau des effectifs. On doit retrouver le tableau de l'ACP.

  - Interprétation : L'analyse des correspondances multiples est une méthode analogue à l'ACP appliquée à des variables qualitatives. On peut faire une ACP sur les individus ou sur les variables du tableau disjonctif complet. En ligne, on analyse les ressemblances entre individus au regard des modalités qu'ils possèdent. On va étudier la variabilité des individus. Pour ce faire, on va extraire les principales dimensions de variabilité avant de les décrire par les modalités. On obtiendra des dimensions décrivant chaque variable initiale avec une hiérarchisation des modalités. En colonne, l'ACM va étudier les associations entre variables et fournira une visualisation d'ensemble d'association entre les modalités. On obtiendra des variables synthétisant au mieux les variables.
    Une proximité des modalités sur le plan traduit une liaison forte entre ces modalités. On peut appeler profil pour simplifier cet ensemble de modalités. Une proximité des individus avec ces profils montre les modalités qui décrivent le mieux ces individus.
    - Premier axe : oppose des modalités et des individus ("effet taille").
    - Second axe : Éloigne les individus remarquables et les modalités rares.

    Le nuage des variables donne une visualisation des variables liées dans chacun des axes. Les liaisons fortes ont de forts coefficients et donc les variables sont éloignées du centre. Les variables proches ont une intensité de liaison proche.

- AFC (Analyse Factorielle des Correspondances)
  - Input : tableau de contingence. On travaille sur un tableau de correspondance entre deux variables qualitatives avec plusieurs modalités. Les valeurs Xij sont le nombre d'occurrence des individus i pour la variable j. C'est un cas particulier de l'ACM où l'on se limite à deux variables.
  - Output : idem que précédemment

  - Transformation : chaque ligne est affectée d'une masse qui est sa somme marginale, le tableau étudié est le tableau des profils des lignes, ce qui permet de représenter dans le même espace à la fois les deux nuages de points associés aux lignes et aux colonnes du tableau de données; elle est par ailleurs très agréablement complétée par des outils de classification ascendante hiérarchique (CAH) qui permettent d'apporter des visions complémentaires, en particulier en construisant des arbres de classification des lignes ou des colonnes.
  - Remarque importante : distorsion de la distance possible dans le cas de modalité rare (distance tellement grande que les distances plus petites vont être cachées).
  - Interprétation : identification des sur/sous-effectifs. proximité = liaison positive = sureffectif / Loin= liaison négative = sous-effectif ; les modalités indépendantes sont au centre ; l'axe horizontale donne une information sur l'intensité de la relation entre les modalités : les plus fortes oppositions sont les plus éloignées.

- AFDM (Analyse Factorielle des Données Mixtes)
  - Input: L'AFDM prend en entrée un tableau contenant à la fois des données quantitatives et des données qualitatives.
  - Transformation : L'algorithme normalise les variables. Le cœur est basé sur l'ACP et l'ACM.
  - Interprétation et output : idem que précédemment

- AFM (Analyse Factorielle Multiple)
  - Input : Tableau avec des variables mixtes et structurées en groupe (GA, GB, GC dans l'illustration ci-jointe). Les variables peuvent être différentes d'un groupe à l'autre et être mixte dans un même groupe. Le cœur est basé sur l'ACP et sur l'ACM.
  - Interprétation et output : idem que précédemment

	GA	GB	GC
Données	xxxxxxxxxx	xxxxx	xxx
i	xxxxxxxxxx	xxxxx	xxx
	xxxxxxxxxx	xxxxx	xxx
	xxxxxxxxxx	xxxxx	xxx
	xxxxxxxxxx	xxxxx	xxx
	xxxxxxxxxx	xxxxx	xxx

- AFD (Analyse Factorielle Discriminante)
  - Input : Un tableau de variables quantitatives avec une variable qualitative à expliquer l'AFD donnera des classes. En représentation on peut faire un nuage de points. Le nombre de classes est connu ou supposé à priori.

  - Interprétation et output : On mesure la qualité de la séparation des groupes mathématiquement et on établit les règles de séparation en rapport avec l'expertise métier.

1.3.2.4 Validation avec le métier

A ce stade, on peut avoir répondu au cahier des charges et présenter les résultats. Une validation est toujours utile avec le métier et d'autant plus si les résultats sont surprenants. Le besoin d'aller plus loin sera guidé par la problématique. Une validation avec le métier est tout aussi importante à ce stade pour éviter d'aller trop loin en cas de soucis sur cette première partie. Dans le planning projet, il faudra positionner un jalon ici.

2 Phase de modélisation

Dans la phase exploratoire on s'est assuré de comprendre de manière approfondie le problème et les données. Dans la phase de modélisation les choix (méthodes, transformations complémentaires, etc.) qui s'imposeront doivent être guidés par le problème à résoudre.

2.1 Généralités sur la modélisation

2.1.1 Généralités sur les modèles

On veut bâtir un modèle descriptif ou prédictif en vue d'un éventuel usage en production. La démarche consiste à construire un modèle probabiliste généralisable en s'appuyant sur le principe d'apprentissage automatique. La méthode sera choisie en fonction du problème à résoudre, du type de modèle attendu mais aussi de critères qualitatifs et des contraintes d'usage en environnement de production. Nous repartons des découvertes effectuées sur notre jeu de données pour sélectionner les méthodes appropriées. Certaines méthodes sont spécialisées dans un type de donnée tandis que d'autres sont plus flexibles. Définissons d'abord l'objet de notre quête.

2.1.1.1 Qu'est-ce qu'un modèle ?

Un modèle est construit par un programme capable de généraliser à partir d'une base d'exemple en s'appuyant sur le principe d'apprentissage. C'est une fonction ou une règle qui "calcule" une valeur en sortie à partir de valeurs en entrée. Pour rendre un peu plus concret le concept de fonction, nous pouvons cités quelques exemples de modèle :

- Dans le cas d'une classification, le modèle est une frontière séparant les classes. Cette frontière prend de nombreuses formes. Dans le cas d'école y=f(x) il s'agit d'une fonction affine simple traduisant une relation linéaire. Elle peut être une fonction quadratique un peu plus complexe. La complexité croit avec le nombre de dimension et la relation n'est plus linéaire.
- S'il s'agit d'un problème de régression alors le modèle est une combinaison linéaire de variables explicatives. Dans sa forme la plus simple c'est également une fonction affine. De manière générale, on obtiendra un vecteur de poids de chaque variable traduisant le pouvoir d'influence de chacune sur la valeur en sortie.
- En clustering, s'agissant d'une approche basée sur des centres et des distances nous obtenons des paramètres valorisés représentants ces centres et ces distances (ex: matrice de similarité).
- Pour les associations, c'est un modèle à base de règles logiques
- Dans le cas des réseaux de neurone, le modèle est basé sur une ou plusieurs fonctions mathématiques appliquées par les nœuds de chaque couche (une fonction par couche).

Le polymorphisme des modèles est dû à la diversité des méthodes de modélisation.

2.1.1.2 Qu'est-ce qu'un bon modèle ?

La qualité d'un modèle prédictif se juge à la précision de la prédiction. On mesure donc le taux d'erreur. Celui-ci devra être le plus bas possible. On appréciera également la stabilité du modèle. Il faut qu'il résiste le plus possible aux valeurs manquantes, aberrantes ou même aux valeurs extrêmes. Il faut aussi éviter le sur-mesure par rapport à l'échantillon d'apprentissage. On considère aussi un modèle simple comme un bon modèle. Si le modèle est trop complexe, l'interprétation devient difficile. Les règles du modèle doivent être lisibles en particulier si l'explication du modèle fait partie du besoin. Si le modèle est issue d'une méthode boite noire, on ne s'attendra pas à une lecture explicite des règles du modèle en revanche la performance temps de calcul et qualité de la prédiction seront les critères d'évaluation de la qualité du modèle.

- Le rasoir d'Occam

« *Nous ne devons admettre comme causes des choses de la nature au-delà de ce qui est à la fois vrai et suffisant à en expliquer l'apparence* » (Isaac Newton). Ce principe dit que le meilleur modèle est le plus simple. En statistique, on l'exprime au travers du dilemme biais-variance (Trade-off en anglais).

- La bonne généralisation ou dilemme biais-variance

En apprentissage automatique, le dilemme (ou compromis) biais–variance est le problème de minimiser simultanément deux sources d'erreurs qui empêchent les algorithmes d'apprentissage supervisé de généraliser au-delà de leur échantillon d'apprentissage.

- Sous-apprentissage :

Le biais est l'erreur provenant d'hypothèses erronées dans l'algorithme d'apprentissage. Un biais élevé peut être lié à un algorithme qui manque de relations pertinentes entre les données en entrée et les sorties prévues.

- Sur-apprentissage :

La variance est l'erreur due à la sensibilité aux petites fluctuations de l'échantillon d'apprentissage. Une variance élevée peut entraîner un sur-apprentissage, c'est-à-dire modéliser le bruit aléatoire des données d'apprentissage plutôt que les sorties prévues.

2.1.2 Principe d'apprentissage

2.1.2.1 Le principe de base

L'algorithme d'apprentissage se résume ainsi :

1. Echantillonnage : train/validation
2. Estimation du modèle : calcul du meilleur modèle pour la méthode sélectionnée à l'aide de la fonction de test choisie

Le principe consiste à séparer le tableau (la population) en deux parties selon des techniques plus ou moins avancées afin de bâtir un modèle sur une partie (échantillon d'apprentissage) et de tester ce modèle sur la seconde partie (échantillon de test). En pratique, on va chercher à estimer un ou plusieurs paramètres à partir du ou des paramètres observés sur les échantillons. On itère autant de fois que nécessaire pour obtenir un modèle satisfaisant une mesure statistique.

2.1.2.2 Les différents types d'apprentissage

- Apprentissage supervisé: les prédictions de l'échantillon d'entrainement sont connus- le modèle est bâti sur une base d'exemple « étiquetés » ;
- Apprentissage non-supervisé : les réponses de l'échantillon d'entrainement sont inconnus – la base d'exemple servant à l'apprentissage n'est pas « étiquetée » ;
- Apprentissage par renforcement : le résultat de la prédiction donne une valeur de retour à l'algorithme d'apprentissage qui en tient compte – on parle de récompense ou de poids attribué à la décision. Cette rétro propagation permet de produire un modèle qui s'adapte aux nouvelle données ;
- Apprentissage profond : apprentissage par superposition de couche de calculateurs cachée implémentant une fonction mathématiques pour évaluer la sortie à chaque couche jusqu'à la valeur finale ;

2.1.3 Élaborer un modèle

Les étapes de la modélisation sont ici décortiquées pour aborder les notions théoriques. En pratique, les étapes sont imbriquées, se succèdent ou s' exécutent en parallèle dans des programmes.

2.1.3.1 Démarche de modélisation

A) Définition de l'objectif d'apprentissage automatique
B) Préparation complémentaires des données (optionnel : selon l'objectif et la méthode)
   1. Agréger – Discrétiser – Décorréler
   2. Identifier valeurs remarquables
   3. Sélectionner les variables du descripteur/prédicteur
C) Echantillonnage en 3 ensembles (Apprentissage/Evaluation/Sélection) - Redresser les échantillons (pour une classification c'est indispensable)
D) Choix des algorithmes pour créer les modèles appropriés à la problématique
   1. Pour chaque algorithme
      i. Répéter plusieurs fois
         a) Estimer un modèle sur l'échantillon d'apprentissage - Choix d'hyperparamètre
         b) Evaluer ce modèle sur l'échantillon d'évaluation - Choix de la fonction de test adaptée à l'algorithme
      ii. Comparer les modèles
      iii. Retenir le meilleur modèle pour un algorithme
   2. Retenir les meilleurs modèles (1 par algorithme)
E) Sélection d'un modèle parmi ceux issues de l'étape D
      i. Choix de la stratégie de test
      ii. Comparaison des performances des modèles sur l'échantillon de Sélection
      iii. Retenir le meilleur modèle
F) Optimisation du meilleur modèle
G) Restitution - Déploiement

L'étape D constitue l'étape de création du modèle. Elle se matérialise dans un programme qui détermine les meilleurs paramètres :
- d'un système relationnel entre les variables explicatives et variable à expliquer dans le cas d'un modèle prédictif
- d'un système décrivant les données dans le cas d'un modèle descriptif

2.1.3.2 Echantillonner pour l'apprentissage

L'objectif consiste à construire un échantillon sur lequel l'algorithme va apprendre. On cherche à avoir une qualité de modèle qui soit équivalente à celui que l'on aurait obtenu avec la totalité de la population disponible dans le jeu de données. De manière générale, il est plus intéressant de travailler avec un échantillon suffisamment grand pour que ces techniques présentent un intérêt. Le schéma d'échantillonnage doit respecter le même schéma que pour la formation du jeu de données (stratification ou tirage par grappes). Il existe différents algorithmes sur lesquels il faut se pencher et expérimenter en fonction de la méthode d'apprentissage qui est envisagée. Un modèle sera bâti sur plusieurs échantillons d'apprentissage partiellement ou totalement indépendants. Des techniques de ré-échantillonnages sont intégrées dans les algorithmes pour l'apprentissage du modèle comme le bootstrap.
Dans ce cas, un échantillon est construit à partir d'un tirage au sort avec remise répété N fois. Cette technique introduit de l'information redondante lors de la construction du modèle. Chaque échantillon de taille identique, mais différent, sert à construire un estimateur simple. Puis un dernier sert à construire un estimateur agrégé plus robuste que chaque estimateur pris individuellement.

2.1.3.3 Estimer un modèle – Choix des hyperparamètres

La méthode consiste à produire une série de paramètres en fonction d'une série d'hyperparamètres. Les réglages à effectuer pendant l'entrainement du modèle portent essentiellement sur les

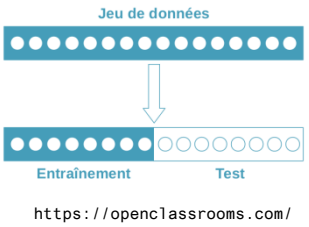
hyperparamètres. Pour estimer un modèle on doit définir un paramètre de complexité : nombre de variables, de voisins, de feuilles, de neurones, durée de l'apprentissage, largeur de fenêtre...; Les paramètres sont estimés à partir des échantillons d'apprentissage. L'apprentissage va modifier les paramètres selon une fonction de coût à minimiser. Pour savoir dans quel sens modifier les valeurs, la majorité des algorithmes s'appuient sur le principe de la descente de gradient. A ce stade on vient de définir un nombre fini de modèles qu'il faut évaluer. L'objectif consistera à retenir les paramètres qui donnent un bon modèle. Cette recherche est effectuée grâce à différentes approches de test.

2.1.3.4 Evaluer un modèle avec l'échantillon de test

C'est un problème d'optimisation des performances du modèle. Pour évaluer les paramètres d'une méthode la technique consiste à mesurer la performance de ces ensembles de paramètres avec une stratégie de validation (échantillon de validation, validation croisée ou bootstrap). On évalue la qualité de la prédiction en apprentissage supervisé (en non supervisé, c'est l'exactitude du modèle qui est évaluée). On utilise une fonction qui calcul l'erreur des modèles issues des échantillons d'apprentissage à l'aide d'une mesure d'erreur statistique (on peut aussi estimer le biais ou l'intervalle de confiance). Les algorithmes implémentant ces techniques proposent de nombreux tests : Erreur quadratique moyenne, erreur absolue moyenne, R^2, etc.

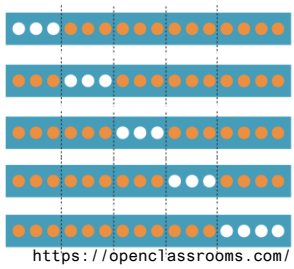
- Echantillonnage de validation :

On utilise une partie du jeu de données n'ayant pas servi à l'apprentissage. Un seul modèle est produit (pas d'itération ni de comparaison de modèle dans ce cas de figure).



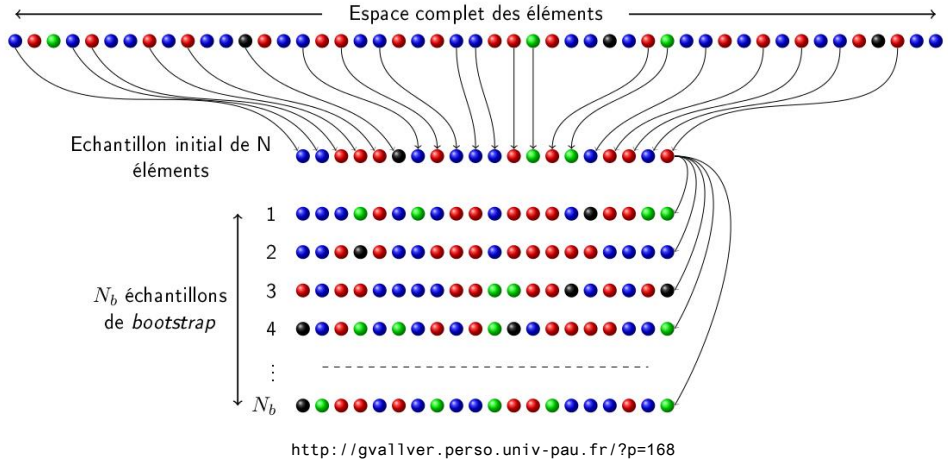
- Validation croisée (cross-validation en anglais) :

C'est une stratégie qui permet d'utiliser le jeu de données entier pour l'apprentissage et pour la validation en répétant l'échantillonnage de base pour obtenir k sous-ensembles (k-fold) du jeu de données complet. On compare plusieurs modèles construits et testés sur des échantillons différents. Dans le cas d'un problème de classification le procédé de stratification permet de redresser les échantillons. Une variante consiste à ne garder pour le test qu'un seul individu (Leave-One-Out).



- Bootstrap :

Cette technique permet d'évaluer la sensibilité des paramètres du modèle sur un échantillon particulier en testant des valeurs sur différents sous-échantillons. Le meilleur modèle est élu par vote majoritaire.



Quelques soit la stratégie de test employée on va chercher à interpréter des mesures afin de juger de de la qualité du modèle. L'ajustement du modèle ou la capacité de prédiction du modèle sont communément évalués. Dans le cas paramétrique, on se rapproche d'une loi de probabilité et l'évaluation repose sur la technique d'optimisation du maximum de vraisemblance. Dans le cas non paramétrique, on ne fait aucune hypothèse sur l'appartenance d'une loi connue et l'évaluation se fait à l'aide de la fonction de répartition empirique. On peut comparer les performances des sous-modèles grâce à des graphiques (Exemple : Aire sous la courbe ROC).

2.1.3.5 Agrégation de modèles

Pour élire le meilleur modèle parmi les sous modèles il existe des méthodes d'optimisation. A la base des méthodes d'ensemble ou agrégation de modèle il y a la technique d'échantillonnage du bootstrap. Dans les méthodes d'ensemble telles que le bagging ou le boosting, la sélection de modèle est intégrée à l'algorithme d'apprentissage selon deux approches distinctes :

- Bagging (Bootstrap AGGregatING)

C'est le même principe d'échantillonnage que le bootstrap (tirage avec remise). Chaque modèle est construit de manière indépendante et en parallèle. La particularité réside dans la méthode d'agrégation des estimateurs faibles. En régression, ce sera la moyenne des poids et en classification ce sera comme pour le bootstrap un vote majoritaire.

- Boosting

L'échantillon d'apprentissage est modifié à chaque itération. Chaque modèle est construit de façon séquentielle en tentant de faire mieux que le précédant modèle. L'algorithme va procéder à des pondérations pour pénaliser ou favoriser un modèle à l'aide d'un nombre très élevé de mini arbre binaire simple (une racine et deux feuilles).

- Stacking

L'agrégation des modèles est effectuée par apprentissage statistique soit à l'aide de méthodes bayésiennes soit à l'aide de réseaux de neurones.

2.1.3.6 Sélectionner et optimiser un modèle

Dans la démarche de modélisation on établit un modèle optimal pour au moins deux procédures différentes appliquées au jeu de données. On évalue les performances obtenues par des modèles issues des deux procédures sur un échantillon qui n'a pas servi à l'élaboration des modèles ni à leur validation. Pour comparer les performances des modèles issues des différentes méthodes on peut utiliser des outils comme la matrice de confusion dans le cas d'un problème de classification et l'aire sous la courbe ROC dans la majorité des cas. En comparant les résultats on sélectionne le modèle issue de la meilleure méthode. On peut chercher à optimiser le modèle sélectionné à nouveau selon la même méthode en testant avec de nouvelles données ou en modifiant les hyperparamètres. Il existe des outils qui permettent d'industrialiser la recherche de méthodes et d'optimisation d'hyperparamètre de manière intelligente pour être le moins gourmand possible en calcul. Les options du programme étant fixées par la main du dataminer, son expérience l'aidera à privilégier les zones de recherches les plus fructueuses.

2.2 Les différentes classes de méthodes

2.2.1 Réduction de dimension

Le principe est de réduire le nombre de variables souvent en vue de diminuer la complexité des programmes. On retrouve les méthodes d'analyse factorielle bien adaptées aux problèmes linéaires. Les méthodes à noyaux traitent les problèmes non-linéaires. Quelques algorithmes sont cités ici : ACP, ACM, AFC, MDA etc.

2.2.2 Régression

En statistique la régression permet d'analyser la relation d'un variable par rapport à une ou plusieurs autres. En modélisation par apprentissage on distingue les problèmes de prévision d'une quantité des problèmes de classement dans des catégories (on parle alors de classification). Le modèle le plus connue est la régression linéaire pour les variables quantitatives. Pour les variables qualitatives la régression logistique est adaptée. Lorsque le modèle n'est pas linéaire, il existe des méthodes permettant de déterminer un modèle prédictif sans connaître la forme de la fonction qui lie les variables entre elles. Si la forme de la fonction est connue (cas de la régression paramétrique), on cherchera à minimiser l'erreur d'approximation avec la méthode des moindres carrées. Si la forme est inconnue (régression non paramétrique), on cherchera à avoir un nuage de points qui présente des variations plus stable. On emploi des techniques de lissage. Il en existe plusieurs :

- par addition de fonction de chaque variable + une erreur
- par plusieurs régression locales (par spline ou polynomiale) qui donne un préditeur par morceau
- par la méthode des noyaux
- par projection qui permet de linéariser le problème en transformant le nuage de point dans un nouvel espace. Le prédicteur est défini dans ce nouvel espace.

2.2.3 Arbre de décision

Un arbre est un graphe avec un nœud racine, des nœuds internes et des nœuds terminaux appelés « feuilles ». Un arbre représente une séquence de décision pour prédire un résultat. Les résultats se situent dans les feuilles. Le résultat peut être une valeur qualitative ou quantitative. Dans sa construction on cherche à maximiser la ressemblance entre les individus au sein des nœuds internes. La ressemblance dépend de l'objectif fixé au départ. Les nœuds les plus hauts testent en premier les variables les plus discriminantes. Le modèle est l'arbre de décision. Il est construit de manière récursive en commençant avec tous les individus dans le nœud racine puis en divisant (split) les individus en groupes les plus différents possibles. Dans le cas des problèmes non-linéaires les arbres peuvent être très larges avec un individu par feuille. Le nœud les plus bas sont généralement très dépendants de l'échantillon. Les arbres sont sujet au sur apprentissage. Par conséquent, il est important d'élaguer l'arbre et de rester à un niveau généraliste. Dans le cas de données manquantes, des algorithmes effectuent des remplacements avec un nœud qui a donné le même split. Quelques algorithmes sont cités ici : CART, C4 .5

2.2.4 Clustering

Les méthodes de clustering visent à regrouper les données en paquets homogènes de sorte que les individus d'un groupe soient les plus semblables et les individus d'un groupe à l'autre soient le plus éloignés possible. Dans tous les algorithmes la notion de base est une mesure de distance, par exemple :

- Distance euclidienne : C'est la distance géométrique entre deux points
- Distance de Mahalanobis : C'est une mesure, basée sur la variance et la corrélation entre variable, qui détermine la similarité entre individus

On cherche à maximiser la distance interclasse et à minimiser la distance intraclasse. C'est un problème très complexe. Il existe plusieurs approches pour trouver la solution optimale dans un temps optimal :

- par partitionnement en fixant le nombre de cluster à priori
- par classification hiérarchique en fixant le nombre de cluster à postériori
- par mesure de la densité
- par calcul probabiliste
- à l'aide de réseau de neurones

La qualité du résultat dépend de la mesure de similarité des individus dans le groupe et de la mesure de la dissemblance entre les groupes. Quelques algorithmes sont cités ici : K-moyennes (partitionnement), DBSCAN (densité), SVM (réseau de neurones), CAH (Classification Ascendante Hiérarchique), Carte auto adaptative (réseau de neurone), etc.

2.2.5 Ensemble

Combinaison de méthodes pour améliorer les performances globales en associant des estimateurs dits faibles. Il existe deux grandes familles : les méthodes séquentielles et les méthodes parallèles.

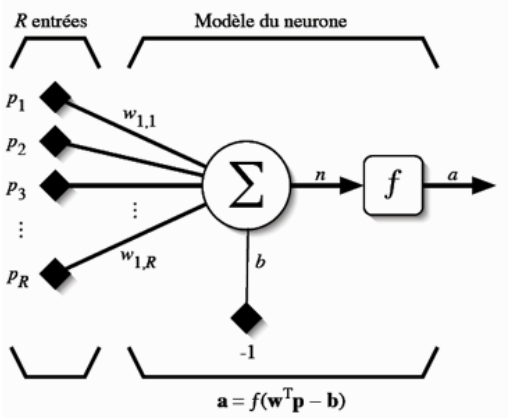
- Parallèle : Ces méthodes génèrent des estimateurs faibles en parallèle. L'apprentissage de chaque estimateur s'effectue sur un échantillon élaboré selon la technique de bootstrap : Bagging, RandomForest.
- Séquentielle : Ces méthodes génèrent des estimateurs faibles les uns à la suite des autres et non pas en parallèle. Chaque itération améliore le précédent par système de boosting en s'appuyant sur le jeu de donnée du précèdent estimateur et en tenant compte de l'erreur précédente : Boosting, Gradient boosting.

2.2.6 Système de règles

Ces méthodes utilisent les notions de support et de confiance pour déterminer la pertinence des règles associatives. Les règles sont générées en fonction de ces valeurs. Quelques algorithmes sont cités ici : Apriori, Eclat, FP-Growth

2.2.7 Réseau de neurones

Inspiré des neurones biologiques, un neurone artificiel se matérialise par une transformation mathématique qui prend R valeurs en entrée et donne une valeur a en sortie. La transformation effectuée est l'application d'une fonction de transfert f à la somme pondérée des entrées (appelée n) moins le biais (b) du neurone.



http://informatique.coursgratuits.net/methodes-numeriques/reseaux-de-neurones-formels.php

Les coefficients w_i représentent le vecteurs poids du neurone. La fonction de transfert peut prendre une forme plus ou moins simple : seuil, linéaire, sigmoïde, tangente hyperbolique, etc.
Un réseau de neurone est un maillage de neurones formels constitué d'au moins une couche où toutes les entrées sont connectées à tous les neurones de la couche d'entrée. Un réseau plus dense consiste à ajouter des couches supplémentaires entre la couche d'entrée et la sortie sur le même schéma. La fonction sera adaptée en fonction de la problématique à résoudre (classification binaire ou multi classes / régression). L'apprentissage neuronal consiste à modifier les valeurs des vecteurs poids et le biais. Un hyperparamètre délicat à régler est le taux d'apprentissage. On peut définir aussi le nombre de couches. Plus il y a de poids plus il faut de données pour éviter le sur-apprentissage. Quelques algorithmes sont cités ici : Perceptron, Perceptron multicouches, Back-Propagation, etc.

2.2.8 Apprentissage profond

Les méthodes d'apprentissage profond reposent sur une architecture élaborée de réseau de neurones. Les couches spécialisées dans une tache se succèdent (réception, compression, correction, estimation de poids, calcul de la perte). Ces méthodes sont applicables aux modèles non-linéaires. Quelques algorithmes sont cités ici : Réseau de neurones convolutifs, etc.

2.2.9 Régularisation

Ces méthodes apportent des solutions à la malédiction de la dimension. Plus le nombre de variables augmente plus l'espace de représentation des données est grand. Cela pose des problèmes de représentation des données et les mesures statistiques ne résistent pas à cet agrandissement. L'interprétation des écarts et les tests de significativité ne sont plus valables. La réduction de dimension est alors intégrer à la méthode d'analyse de donnée. Quelques algorithmes sont cités ici : Ridge, Lasso, etc.

2.2.10 Méthodes topologiques

Ces méthodes tiennent compte de la distance topologique entre individus. Si cela a du sens d'en tenir compte dans la problématique métier c'est un bon point pour les choisir. Quelques algorithmes sont cités ici : K plus proches voisins, Carte auto adaptative de Kohonen (également considéré comme une méthode neuronale), etc.

2.2.11 Bayésien

Ces méthodes se basent sur le théorème de Bayes utilisé en statistique inférentielle pour déterminer la probabilité de cause. Ce théorème est utilisé pour mettre à jour les paramètres du modèle probabiliste. Les données peuvent être bruitées, la méthode réagira bien. Il en existe pour la prédiction de variable qualitative ou quantitative. Le classifieur bayésien est présenté comme l'étalon pour évaluer la performance des autres algorithmes. Le résultat obtenu en sortie est un tableau de croisés de variables explicatives avec la variable à expliquer. Quelques algorithmes sont cités ici : Naive Bayes, Réseau bayésien, Gaussien Naive Bayes, etc.

2.3 Quelques exemple de choix de méthodes selon l'objectif

Le travail du dataminer consiste à choisir quelques méthodes répondant au besoin. L'implémentation des algorithmes exigent de définir des hyperparamètres. On trouve ici quelques exemples de choix de méthodes selon l'objectif de l'étude.

2.3.1 Décrire

On veut découvrir le meilleur descripteur en résumant les données. On se trouve dans la situation suivante : pas d'hypothèse préalable, pas de sélection de variable à priori, pas de valeur cible à prédire ou à expliquer. On résume les variables avec l'analyse factorielle. Grâce à cette technique on réduit le nombre de variable en remplaçant les variables corrélées par une nouvelle variable qui est une combinaison linéaire des variables corrélées. Cette technique ne fonctionne pas toujours en particulier si le problème est non linéaire. Si on souhaite ensuite faire un modèle prédictif et que l'on a trop de variables on peut chercher à utiliser des méthodes de prédiction qui intègrent automatiquement la sélection de prédicteurs en régression :

- Ridge : réduit les poids des variables corrélées
- Lasso : poids = 0 pour sélectionner les variables

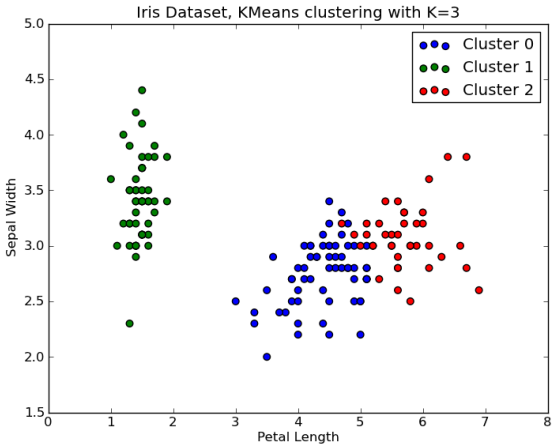
On peut aussi utiliser les méthodes à noyaux pour traiter les problèmes non linéaires. L'astuce consiste à changer d'espace et à appliquer la méthode dans un espace de redescription dans lequel le problème est linéaire. Il existe aussi des méthodes d'ensemble pour la sélection de variables automatique.

2.3.2 Structurer

On veut structurer l'information en regroupant les données sans vouloir sélectionner de variable ni prédire ou expliquer de valeur cible. On peut avoir à formuler des hypothèses. Certaines méthodes peuvent être utilisées pour discrétiser des variables.
Quelques algorithmes sont cités ici :

2.3.2.1 K-Means

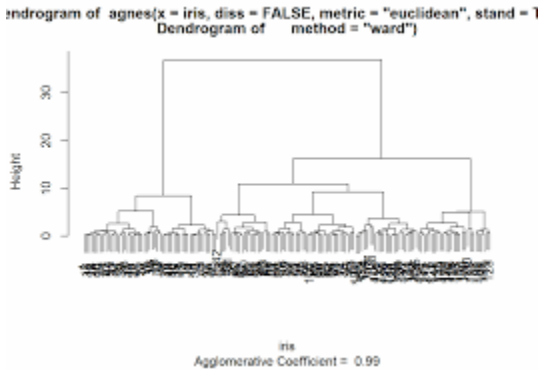
- Hyperparamètre à fixer : nombre de centre moyen - K nombre de partitions
- Approche : partitionnement
- Métrique : distance
- Graphique : nuage de points



https://stackoverflow.com/questions/6645895/calculating-the-percentage-of-variance-measure-for-k-means

2.3.2.2 CAH (Classification Ascendante Hiérarchique)

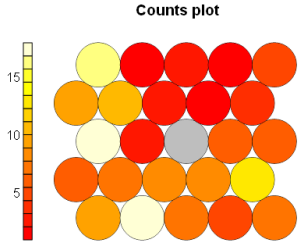
- Hyperparamètre à fixer : aucun à priori
- Approche : hiérarchique
- Métrique : similarité
- Graphique : dendrogramme (permet de choisir à postériori le nombre de classe par coupure)



https://rstudio-pubs-static.s3.amazonaws.com/86360_4b5495d5aefd4821a1981f95774dd2d3.html

2.3.2.3 Kohonen (Carte auto organisée)

- Hyperparamètre à fixer : nombre d'unités cachées
- Approche : connexionniste
- Métrique : distance
- Erreur à minimiser : La somme des carrées des distances
- Graphique : carte topologique



https://help.xlstat.com/customer/fr/portal/articles/2910042-cartes-auto-organisatrices-kohonen-dans-excel?b_id=9283

2.3.3 Prédire

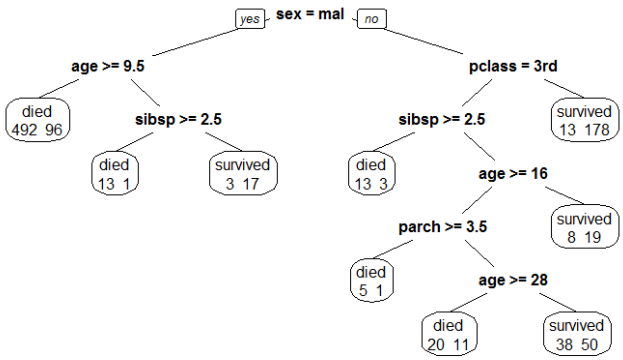
On peut déterminer la classe de méthode pour résoudre cette problématique en regardant la nature de la donnée à prédire : quantitative ou qualitative ?

2.3.3.1 Prédire sans modéliser

- K plus proches voisins (KNN):
  - Prétraitement : Attention au nombre de dimensions car la notion de distance devient peu pertinente dans ce cas. Il faudra donc envisager une réduction de dimension si cela ne présente pas d'impact sur la représentativité des informations ainsi obtenues.
  - Hyperparamètre : Il faut définir le meilleur K pour une meilleure capacité à prédire
  - Approche : topologique
  - Métrique : distance
  - Erreur à minimiser : La somme des carrées des distances
  - Graphique : nuage de points

2.3.3.2 Classification pour prédire une classe

- Arbre de décision CART :
  - Prétraitement : mal adapté aux très grands échantillons
  - Hyperparamètre : Taille minimale d'un nœud, profondeur de l'arbre,
  - Approche : arbre binaire et classification
  - Erreur à minimiser : erreur de prédiction / Elagage (pruning en anglais) de l'arbre au niveau d'erreur minimale
  - Graphique : Arbre de décision



https://lovelyanalytics.com/2016/08/18/un-arbre-de-decision-avec-r/

Illustration 4 : Arbre CART en R avec le Dataset Titanic (Indice de GINI comme critère de sélection)

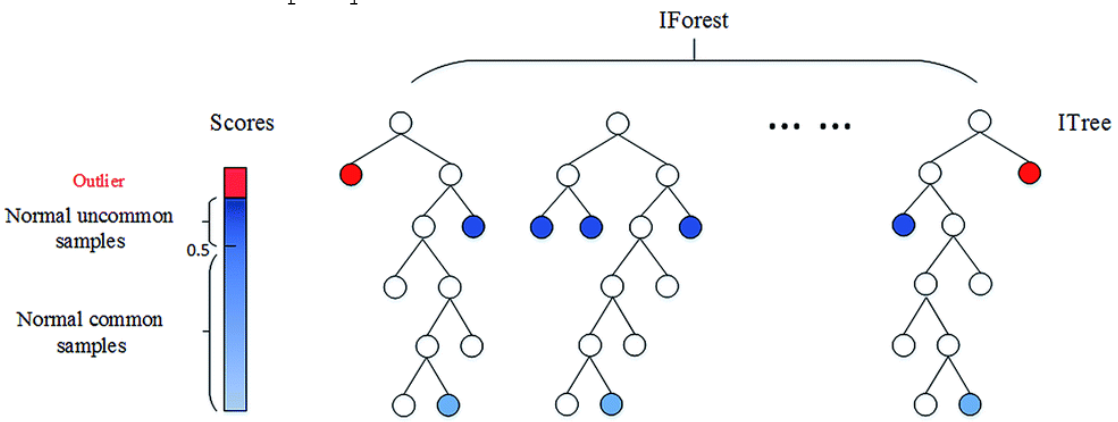
- Alternatives : Régression logistique ou SVM

2.3.3.3 Régression (linéaire ou non-linéaire) pour prédire une valeur numérique

- SGD Regressor :
  - Hyperparamètre : fonction de perte linéaire ou non-linéaire, pénalité
  - Approche : Descente de gradient
  - Métrique : moindres carrés en général
  - Graphique : selon la problématique

2.3.4 Détecter

- Isolation Forest :
  - Hyperparamètre : nombre de cluster, fonction de décision et seuil
  - Approche : Arbre de décision
  - Métrique : profondeur de l'arbre (très petite pour les outliers)
  - Graphique :



http://pubs.rsc.org/en/content/articlelanding/2016/ay/c6ay01574c/unauth#!divAbstract

2.3.5 Associer

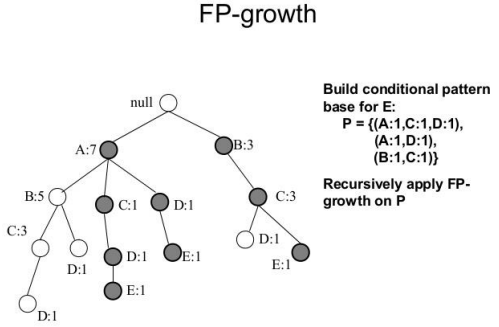
Les algorithmes utilisent la notion de support et de confiance pour générer les règles associatives. Elles sont validées si le support et une confiance tout deux supérieurs à un seuil fixé pour chacun.

2.3.5.1 Apriori

- Hyperparamètre : Seuils support et confiance
- Approche : Calcul des fréquences des ensembles d'items
- Métrique : support et confiance
- Limites : performants sauf s'il existe un nombre important d'ensembles fréquents

2.3.5.2 FP-Growth

- Hyperparamètre : Seuils support et confiance
- Approche : Arbre
- Métrique : support et confiance
- Avantage : Par rapport à l'algorithme Apriori le grand nombre de motif fréquent ne pose pas de problème. La structure compacte de l'arbre permet de tracer dans les différents parcours de l'arbre les ensembles fréquents.



https://www.slideshare.net/wanaezwani/apriori-and-eclat-algorithm-in-association-rule-mining

2.4 Déployer

2.4.1 Restituer les résultats

2.4.1.1 Documenter

- ✓ Rappeler le contexte (Pourquoi) : Décrire en quelques mots la problématique métier
- ✓ Problème (Question) : Rappeler la formulation de la question en problème de fouille de donnée
- ✓ Solution (Réponse) : Décrire succinctement le résultat
- ✓ Découvertes : Lister les découvertes sur les données et sur les méthodes employées (qui ont fonctionné ou pas). Fournir des informations sur les performances
- ✓ Limites: Décrire les limites d'utilisation du modèle avec une indication sur sa qualité
- ✓ Résumé : Ecrire un résumé des trois premiers points pour faciliter la relecture ultérieur du rapport.

2.4.1.2 Présenter

Illustration 5: Les utilisateurs de Power BI disposent de plus de 35 types de visualisations différentes.

Avec des visualisations adaptées aux données (notamment à l'échelle des différentes variables) les possibilités de représentation explosent. La data visualisation devient un art mais surtout (et ce sont les points essentiels à retenir) c'est qu'une bonne représentation des données :

- permet avant tout une bonne interprétation et compréhension des données
- augmente l'impact de la communication des résultats et donc l'adhésion du métier

2.4.1.3 Transposer les résultats en décisions métiers

Un visuel à vocation à représenter les découvertes issues de l'analyse de données mais aussi le modèle issu des algorithmes. Ce dernier ne peut pas être directement exploité. Il faut revenir au problème métier pour transposer les résultats dans le langage métier et les traduire en décision à prendre ou opportunités à saisir. Une connaissance du métier est primordiale à cette étape.

2.4.2 Passage en production

Une fois le modèle établi il doit la plupart du temps être recoder par un développeur sur la plateforme de production pour pouvoir être exploité.

Conclusion

- La démarche de fouille de données peut être généralisée dans un projet d'entreprise. Il peut y avoir plusieurs étapes de modélisation avant d'aboutir à un modèle. On découpe le problème principal en sous problème. On peut formuler un objectif par étape intermédiaire. Un modèle en sortie d'une étape peut être mis en entrée d'une autre étape et ainsi de suite jusqu'à l'obtention d'une solution au problème général. Par exemple, on peut d'abord décrire et sélectionner des variables puis structurer les données et enfin détecter des anomalies ou bien prédire par groupe. Tout dépend du chemin à parcourir entre les données de départ et la solution à mettre en œuvre.
- Il y a des étapes cruciales à ne pas rater comme le choix de la méthode et validation du modèle.
- Le choix de méthode est lié souvent à la tradition ou aux habitudes du secteur qui utilise souvent les mêmes.
- A la construction du planning et pour la répartition des charges, il faut prendre soin de considérer que l'estimation du temps à consacrer aux différentes étapes n'est pas proportionnel à l'importance de l'étape elle-même.
- Dans ce type de projet, les deux principaux écueils à éviter sont les suivants :
  - On n'a pas besoin de faire des hypothèses avant de produire un modèle prédictif. Dans les faits si on ne suppose rien et que l'on applique une des méthodes de prédiction sans se poser de question alors la prédiction sera aussi bonne qu'une prédiction aléatoire ;
  - On peut manipuler un modèle pour obtenir ce que l'on souhaite. Il faut accepter que le résultat ne soit pas celui que l'on attend.

Annexes

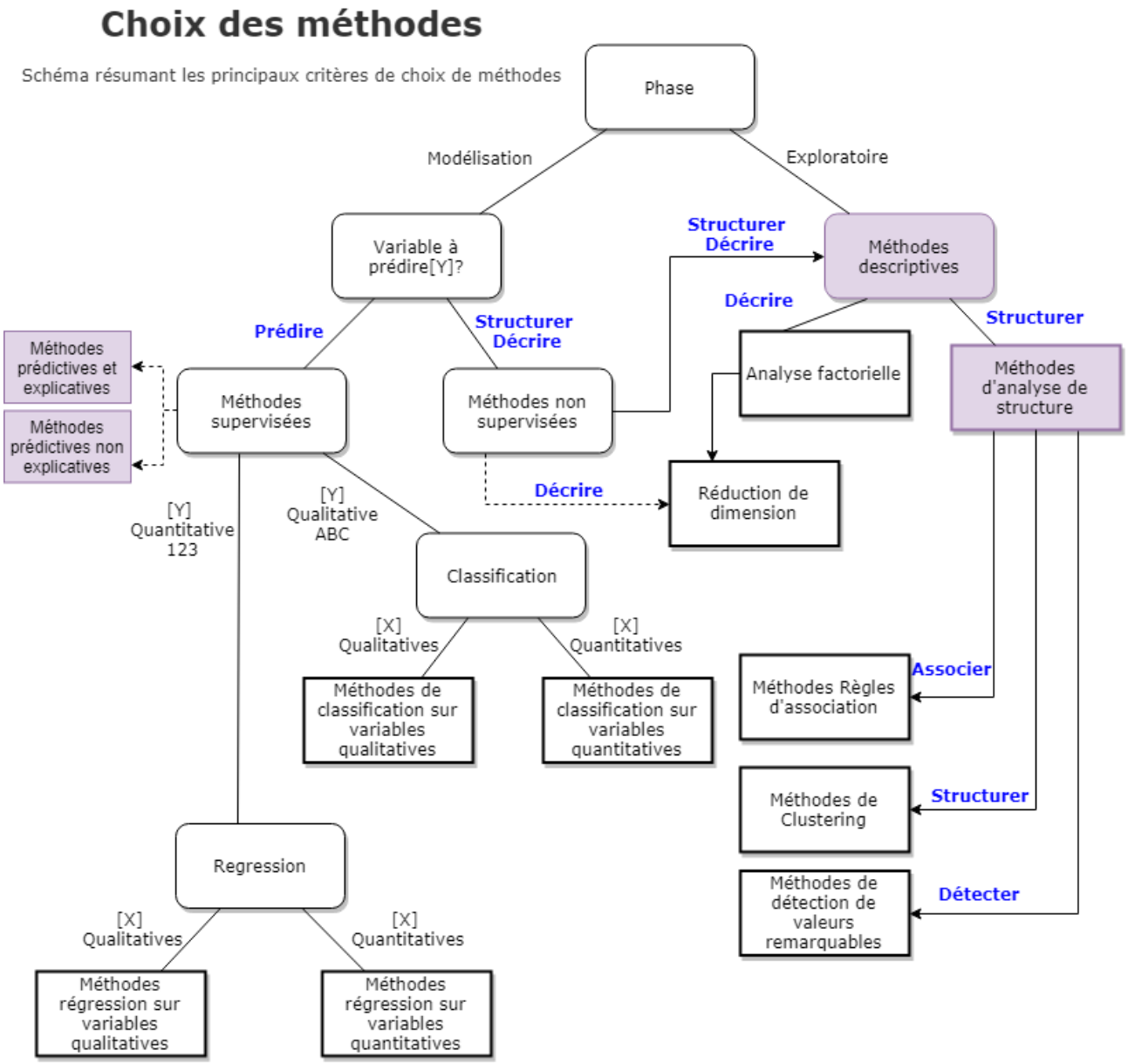


Illustration 6 : Diagramme de sélection de méthodes (très personnel)

Webographie

http://www.cons-dev.org/elearning/stat/multivarie/6-2/6-2.html
http://rb.ec-lille.fr/l/Analyse_de_donnees/Methodologie_L_AFC_pour_les_nuls.pdf

Illustrations