L'ANALYSE DE DONNÉES

L'analyse de données, qu'est-ce que c'est ?

L'accroissement des processus métiers gérés par l'informatique (digitalisation) a fait exploser le volume et la variété des données produites depuis la bulle internet puis mobile. On peut accéder à des données externes telles que les open datas ou les réseaux sociaux. Savoir exploiter ces données est un enjeu important pour les entreprises : d'abord parce qu'il est essentiel de s'assurer de la pertinence de celles-ci ; d'autre part parce que leurs résultats d'analyse fournissent des informations importantes permettant d'améliorer, d'optimiser ou d'anticiper les processus métiers. Les domaines d'application sont nombreux et hétéroclites : Banque/ Finance, Assurance, Médecine, Météorologie, Sociologie. L'objectif consiste, à présent, à avoir des outils qui intègrent plus de données et qui les analysent efficacement.

L'analyse de données est un ensemble de méthodes statistiques appliquées à un jeu de données dans le but d'extraire des informations pertinentes ; on appelle cette extraction fouille de données. Le but est de dégager des tendances, des profils, de détecter des comportements ou de trouver des liens, des règles. Il existe deux grands types d'analyse de données : l'analyse descriptive et l'analyse prédictive.

L'analyse descriptive a pour but de résumer les données en leur assignant une nouvelle représentation, de synthétiser en faisant ressortir ce qui est dissimulé par le volume. On peut classer les individus dans des catégories, trouver les individus les plus proches ou les plus éloignés entre eux ; mais aussi trouver les exceptions ou les cas atypiques. On peut également voir si des variables sont proches, expliquer une variable en fonction des autres ou encore repérer les variables les plus influentes.

L'analyse prédictive consiste à analyser les données actuelles afin de faire des hypothèses sur des comportements futurs. On se sert des données que l'on possède déjà pour extrapoler et deviner le comportement de nouveaux individus mais également l'évolution des individus déjà présents.

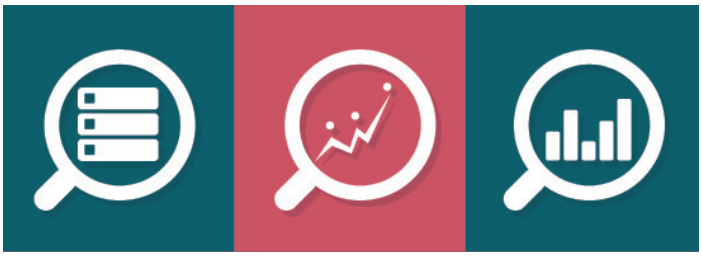
L'analyse de données par étape

Une analyse de données doit suivre une démarche bien particulière. Les métiers de Data Analyst et de Data Scientist jouent chacun un rôle bien spécifique dans cette démarche.

DATA SCIENTIST

Définition de l'information recherchée : spécifier les « problèmes posés ».

Récupération des données : on doit s'assurer de la fiabilité, de la pertinence et de l'intégrité des données à analyser.



Etude des résultats : les résultats obtenus sont ensuite exploités, on cherche à leur donner un sens.

Exploitation des résultats : finalement les résultats de l'analyse permettent de répondre aux « problèmes posés » et contribuent donc à l'aide à la décision.

DATA ANALYST

Nettoyage des données : les données proviennent généralement de divers formats et peuvent être incomplètes ou non structurées. Il faut donc les nettoyer et les organiser.

On les présente généralement sous forme de tableau : une ligne représente un « individu » et une colonne représente une « caractéristique ».

NOM	PRECIPITATION	TEMP.MAX	TEMP.MIN
RENNES	15	37	2
PARIS	23	34	-6
BREST	39	31	0

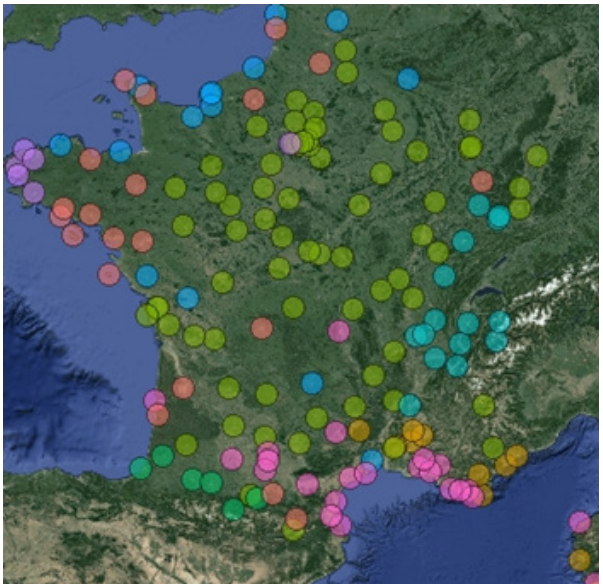
Construction du modèle : la mission consiste ensuite à trouver le ou les bons modèles et les méthodes associés en fonction des données et des informations recherchées. Puis des tests sont effectués pour valider le modèle.

Exemples

Nous allons analyser les données météorologiques de 175 villes françaises sur l'année 2015. Nous avons 5 relevés par mois : Température Maximum Absolue, Température Minimum Absolue, Température Maximum Moyenne, Température Minimum Moyenne, Précipitation. C'est donc un jeu de données de 175 individus et 60 caractéristiques (12 mois x 5 relevés). Ici les données sont fiables, complètes et déjà ordonnées. Le but est simplement de présenter différents modèles descriptifs ou prédictifs mais également la complémentarité entre ces deux types d'analyse.

Analyse Descriptive

Nous commençons donc par décrire nos données grâce à la méthode des k-means qui permet de déterminer des classes d'individus sur un échantillon de données.



K-means :

Nous avons ajusté les différents paramètres de l'algorithme afin d'avoir une pertinence au niveau de la répartition des villes. Grâce aux données météorologiques, l'algorithme a déterminé comme proches des villes qui sont effectivement voisines géographiquement.

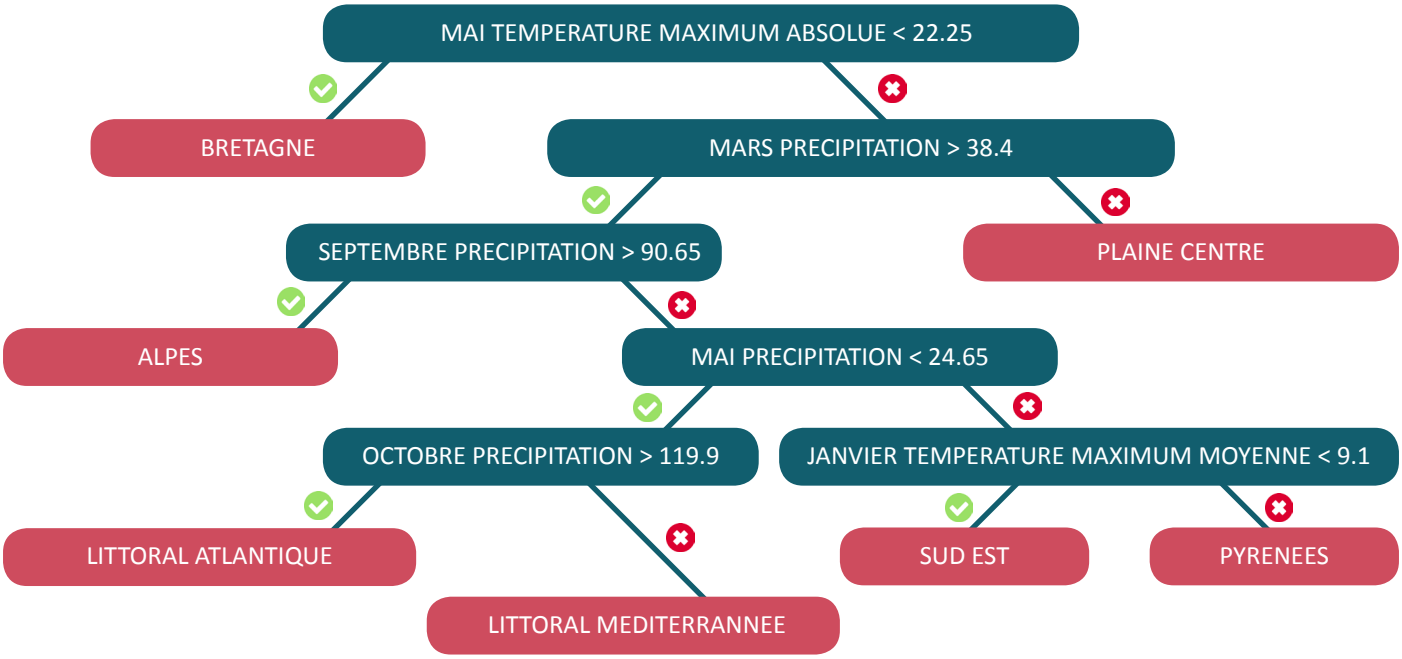
Une couleur est associée à chaque groupe déterminé par l'algorithme. On pourrait nommer ces groupes : Littoral Nord, Bretagne Ouest, Littoral Atlantique, Plaine Centre, Alpes, Pyrénées, Littoral Méditerranée et Sud Est.

L'algorithme de k-means est beaucoup utilisé dans le marketing dans le but de catégoriser les clients pour mieux cibler les démarchages de produit. Le service Etudes Marketing du Crédit Agricole Centre France a par exemple distingué 3 profils : petit consommateur, couple actif et couple boursicoteur.

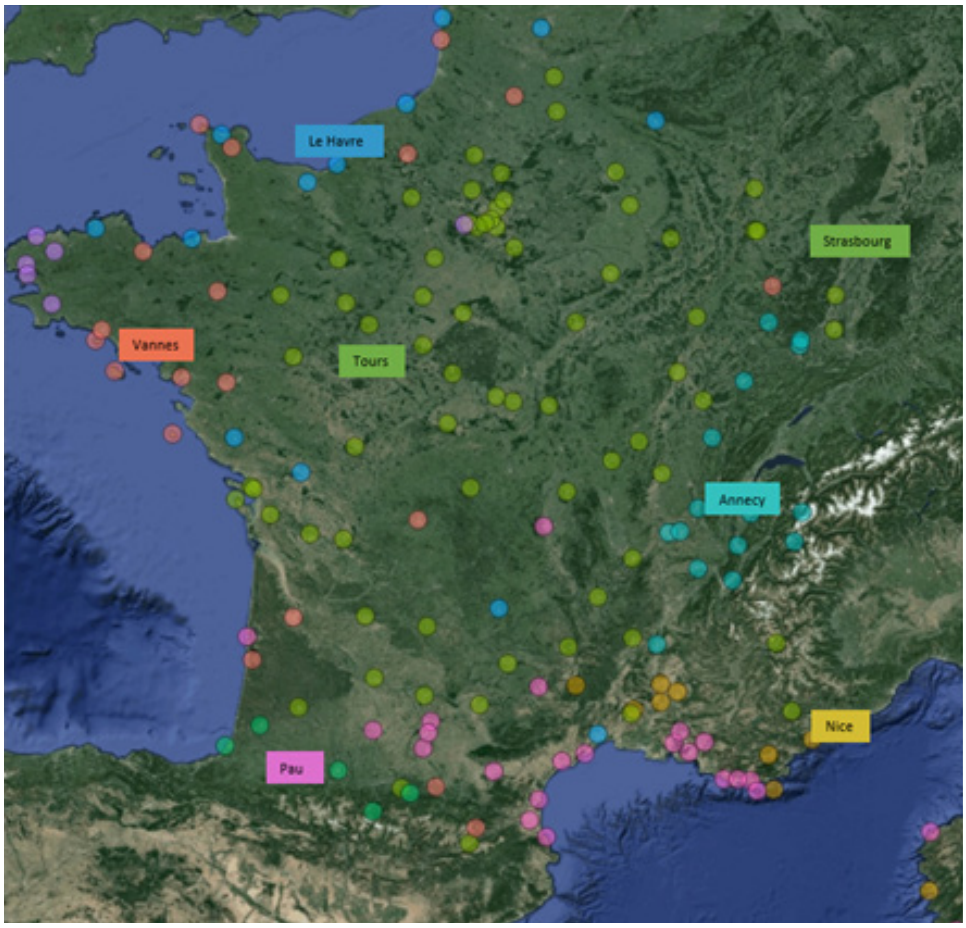
Analyse Prédictive

Nous décidons ensuite de nous servir des résultats obtenus grâce à l'analyse descriptive, à savoir les 8 groupes définis par la méthode des k-means pour appliquer un modèle d'analyse prédictive à nos données. On construit donc un arbre de décision qui détermine en fonction des données météorologiques de la ville le groupe auquel elle appartient.

Arbre de décision :
Le principe est qu'à chaque nœud, on pose une question sur un caractère de l'individu puis on suit la branche qui correspond à la réponse pour arriver à un autre nœud. A la fin de l'arbre, on arrive sur une feuille qui nous donne le groupe auquel l'individu a de très fortes chances d'appartenir. Les caractéristiques et les valeurs de chaque nœud ainsi que l'ordre des nœuds sont choisis automatiquement par l'algorithme.



Nous regardons dans quel groupe l'arbre classe 7 nouvelles villes. Ces villes sont placées sur la carte avec leur nom et la couleur du groupe dans lequel l'arbre les a classées. 6 d'entre elles semblent bien classées ; Pau a été classé en Littoral Méditerranées alors qu'il paraît plus proche du groupe Pyrénées, les deux groupes sont néanmoins voisins.
L'arbre de décision est un des modèles d'analyse prédictive les plus utilisés.



La Faculté de Médecine-Pharmacie de l'Université de Rouen l'utilise par exemple pour détecter le risque d'Embolie pulmonaire sur un sujet âgé.

Enfin, nous décidons d'appliquer un autre modèle d'analyse prédictive.
Nous réalisons donc une régression. Le but est de prédire la Température Maximum Absolue d'un mois en fonction de toutes les données météorologiques du mois précédent.

Régression :
Nous retirons les individus Lyon et St Brieuc de la base avant de construire le modèle que nous testons ensuite avec ces mêmes individus. On peut comparer les valeurs prédites avec les valeurs réelles :

	PREDICTION AVEC MODELE SUR TOUTE LA BASE	VALEUR REELLE
LYON	36.5°C	37.4°C
St-BRIEUC	33°C	27.3°C

Le résultat de la prédiction est assez éloigné de la valeur réelle. Nous décidons donc d'utiliser une nouvelle information : celle du groupe auquel appartient la ville (groupes déterminés par la méthode des k-means). Nous construisons alors un modèle régressif par groupe de ville. On peut à nouveau comparer les valeurs :

	PREDICTION AVEC MODELE SUR TOUTE LA BASE	PREDICTION AVEC MODELE SUR LE GROUPE ALPES	PREDICTION AVEC MODELE SUR LE GROUPE BRETAGNE	VALEUR REELLE
LYON	36.5°C	37.8°C		37.4°C
St-BRIEUC	33°C		28°C	27.3°C

On remarque que le modèle est beaucoup plus précis lorsqu'il est construit sur des villes du même groupe, c'est-à-dire des villes ayant des caractéristiques proches. Cet exemple fait ressortir la complémentarité entre l'analyse descriptive et l'analyse prédictive.

La régression permet d'estimer une donnée numérique. Axa s'en sert pour modéliser la fréquence et le coût moyen du sinistre en assurance automobile du particulier.

Construction du modèle : la fouille des Algorithmes

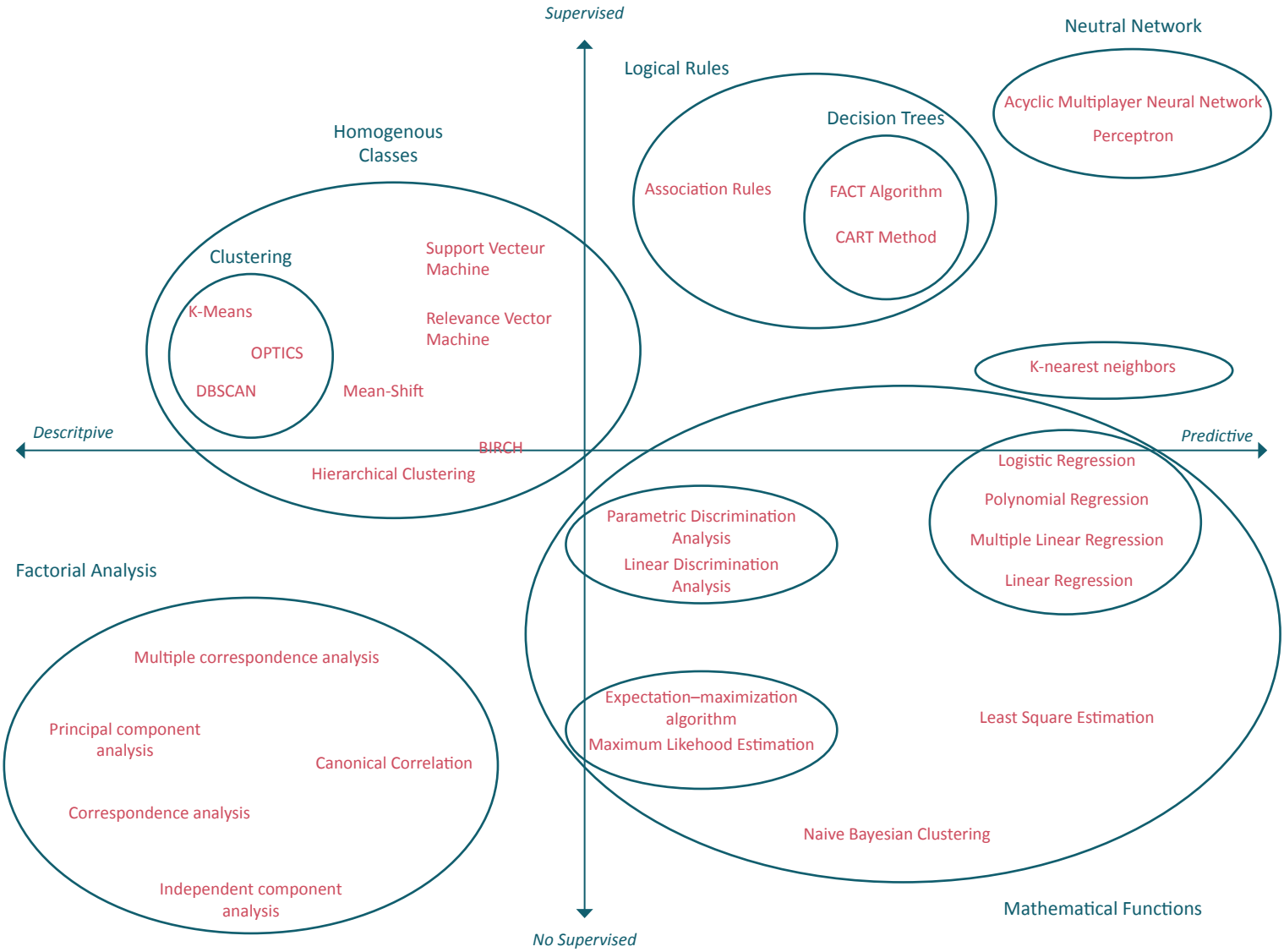
Dans cette partie, nous allons détailler l'étape appelée construction du modèle ; certaines notions peuvent être difficiles à aborder pour des lecteurs non avertis. Le Data Analyst doit trouver la ou les bonnes méthodes pour construire ce modèle et l'appliquer au jeu de données. Pour cela on adopte une démarche appelée « Machine Learning ».

La machine se « nourrit » des données afin d'établir le modèle statistique, cette phase est appelée apprentissage. L'apprentissage peut être supervisé ou non. Pour l'apprentissage supervisé, on dit à la machine quelle donnée doit être expliquée ; pour l'arbre de décision par exemple, les villes sont étiquetées avec leur groupe. L'algorithme k-means en revanche est non supervisé puisqu'il détermine automatiquement des groupes homogènes au sein du jeu de données.

Pour les modèles supervisés, le principe est de découper le jeu de données en deux parties. Les deux premiers tiers des individus, appelés échantillon d'apprentissage servent à construire le modèle. On doit vérifier la pertinence de celui-ci pour pouvoir l'affiner ensuite ; on utilise donc le dernier tiers des données, appelé échantillon test pour comparer le résultat réel avec le résultat donné par le modèle.

Il existe des tests pour mesurer la fiabilité du modèle. Pour les valeurs numériques, on calcule la différence entre la valeur obtenue par le modèle et la valeur réelle. Pour les classifications, on regarde la proportion d'individus mal classés. On essaye donc de réduire ce rapport, cependant si le pourcentage d'erreur est inférieur à 5%, il s'agit de sur-apprentissage ; c'est-à-dire que le modèle « colle » trop à l'échantillon d'apprentissage et qu'il ne saura pas prédire de nouveaux individus. De plus, selon les cas, certaines mauvaises classifications sont plus néfastes que d'autres. Par exemple il vaut mieux qu'un mail « spam » soit classé en « non spam » que l'inverse. Cela doit également être réfléchi par le Data Scientist et réglé par le Data Analyst.

Avant de vous décrire les algorithmes les plus fréquemment rencontrés, le schéma ci-dessous illustre les grandes familles de modèles et de méthodes et leurs répartitions selon les deux axes majeurs que sont : Prédictif/Descriptif et Supervisé/Non Supervisé.



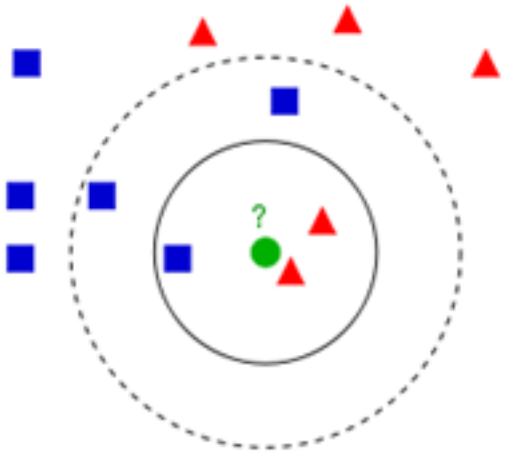
Principal Component Analysis :
L'analyse en composantes principales réduit tous les caractères d'un individu en seulement deux dimensions. L'ensemble des individus peut donc être représenté par un nuage de points sur un plan de projection. On cherche le plan de projection le plus représentatif.

Clustering :
La classification sert à découper une population en groupes les plus homogènes possibles. On peut connaître ou non nos groupes à l'avance. On utilise l'écart entre les individus qu'on calcule à partir de la différence entre leurs caractères. Ainsi deux individus très proches seront dans le même groupe.

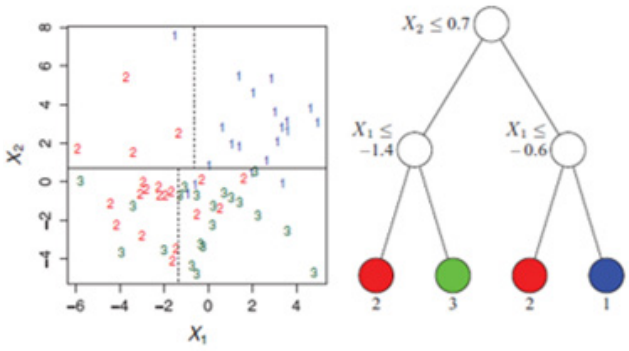
Naive Bayesian Clustering :
La classification bayésienne se base sur la formule de probabilité conditionnelle de Bayes. On se sert des probabilités de l'échantillon d'apprentissage pour déterminer la probabilité qu'un individu appartienne à une classe donnée. On le classera dans le groupe pour lequel il a obtenu la plus forte probabilité.

Association Rules
Le but de la détection d'association est de trouver des règles logiques. C'est très utilisé pour les applications commerciales. Par exemple : lorsque les gens achètent le produit A et le produit B, ils achèteront le produit C avec un indice de confiance. On peut donc prédire les prochains achats en fonction des achats déjà effectués.

Knn :
On détermine le groupe d'un nouvel individu avec un calcul de distance : il appartiendra au groupe de ses plus proches voisins. L'élément vert est classé rouge si on choisit ses 3 plus proches voisins et bleu si c'est 5.

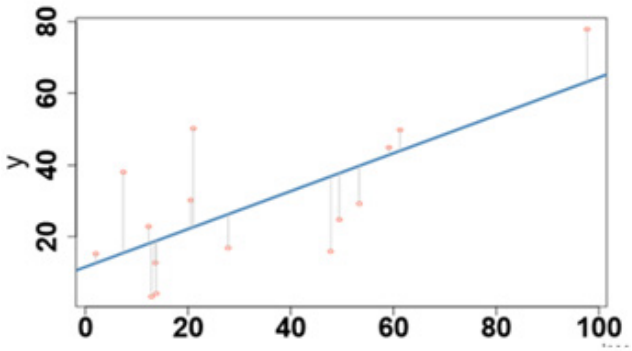


Decision Tree :
L'arbre de décision classe un individu dans un groupe. A chaque nœud on s'intéresse à une variable et en fonction de la valeur de celle-ci, on suit une branche pour arriver à un nouveau nœud, jusqu'à arriver à une feuille qui donne le groupe de l'individu. L'arbre est construit étape par étape : on cherche d'abord pour quelle variable et pour quelle valeur de cette variable le groupe est le mieux partitionné, et ainsi de suite comme on peut le voir sur l'exemple ci-dessous. Pour classer un nouvel individu, il suffit de « suivre » l'arbre.



Regression :
La régression établit la relation mathématique entre une variable et toutes les autres. Les paramètres doivent minimiser l'écart entre la fonction et la véritable variable à prédire de l'échantillon d'apprentissage.

Droite de régression (au plus proche des points)



Le Deep Learning

Le « Deep Learning », littéralement « apprentissage profond », connaît un engouement très rapide. C'est un système d'apprentissage automatique basé sur les réseaux de neurones artificiels. Inspirés du fonctionnement des neurones humains, les réseaux de neurones artificiels sont des unités de calculs élémentaires interconnectées. Grâce aux réseaux de neurones, l'apprentissage fonctionne en strates qui affinent la (re)connaissance étape par étape, ce qui augmente considérablement la faculté d'apprentissage. Le progrès réside dans le fait que ces strates sont très nombreuses, c'est pourquoi on le qualifie d'apprentissage profond. La machine garde en mémoire ces connaissances, ce qui permet ensuite d'augmenter fortement la vitesse des calculs. Elle sait, par exemple, représenter et reconnaître « seule » des schémas dans des données telles que la parole ou l'image.

L'utilisation du Deep Learning en entreprise est naissante mais connaît déjà des résultats très concluants notamment dans le domaine du marketing et de ciblage clientèle. L'entreprise Binatix l'utilise également pour établir sa stratégie de trading.

Les outils

R, SAS, Python ou encore Matlab sont les principaux langages interprétés d'analyse de données. Ils intègrent des données provenant de différents types de fichiers tels que des fichiers textes ou des tableurs, il ne s'agit cependant que de fichiers plats et structurés donc très éloignés du Big Data. On peut aussi faire des appels vers des bases de données. Ils permettent d'exécuter des manipulations et des calculs mathématiques simples sur différents objets. Ils offrent également une très large bibliothèque de fonctions statistiques extensibles avec des packages grâce auxquels on peut faire appel à des algorithmes déjà implémentés. Ces outils présentent donc des lacunes en termes d'accès aux données, d'étude de la qualité des données mais également en visualisation des données.

Il existe également des logiciels d'analyse statistique avec une interface simple tels que Oracle E-Business Suite, Watson (édité par IBM), SAS BI, KNIME, RevolutionR de Revolution Analytics, Tableau ou encore DataIKU où on peut lancer facilement une analyse de données qui construit intuitivement et rapidement des modèles sur des données qui peuvent provenir de n'importe quelle source de données. Contrairement aux langages interprétés, il arrive que l'humain n'ait aucun contrôle sur les opérations : les différents paramètres des algorithmes sont choisis par le logiciel et le Data Scientist ne peut pas ajuster ou valider le modèle.

L'enjeu actuel est d'avoir des outils qui analysent plus de données, plus vite et surtout qui intègrent facilement des données plus hétérogènes et moins structurées ; le but étant de pouvoir analyser en direct un flux continu de données afin de pouvoir exploiter les résultats en temps réel. Certains commencent à apparaître.

Vous voulez approfondir le sujet avec nous ?

Voici les liens où vous pourrez trouver les cas d'usage en entreprises qui ont été cités :

http://math.univ-bpclermont.fr/biblio/rapport/banque/2011/M2_Hammouali_11.pdf
http://www.chu-rouen.fr/page/arbres-de-decision
http://www.ressources-actuarielles.net/C12574E200674F5B/0/EDD93736CF1E054EC1257871004C4C87
http://www.affinio.com/blog/deep-learning-disrupting-social-marketing-and-advertising

Dans le cadre des recherches de la cellule Innovation, Aubay poursuit actuellement plusieurs projets de R&D afin d'approfondir son expertise technique autour de l'Analyse de données et de mettre en œuvre de nouvelles solutions dans des approches techniques novatrices. Si vous êtes intéressé soit pour être tenu au courant de l'avancée de ces travaux et des résultats obtenus soit pour participer avec nous au déroulement de ces travaux, n'hésitez pas à nous contacter (innov-dt@aubay.com).