

## A propos de ce document

Ce document a pour objectif de répondre aux interrogations sur ce qui différencie KXEN dans son approche de l'Analyse de Données et plus particulièrement son approche du Data Mining. Cette introduction ne va pas vous plonger dans un comparatif sur les Mathématiques et l'Architecture Informatique qui différencient KXEN des autres éditeurs de logiciels de Data Mining mais explique les raisons qui nous ont amenées à créer KXEN.

Les décideurs à la recherche d'outils ou de techniques d'Analyse Prédictive qui ont déjà fait le tour de plusieurs fournisseurs en la matière seront intéressés par les propos qui suivent. Cela leur permettra de faire un benchmark de KXEN par rapport à d'autres techniques de Data Mining et de découvrir comment KXEN peut les aider à faire des profits à partir des mathématiques et des dernières avancées dans le domaine du « machine learning » tout en révolutionnant leur approche du Data Mining.

## KXEN est orienté résultats

L'aspect primordial pour comprendre le positionnement de KXEN est de garder à l'esprit que KXEN est orienté **résultats** plutôt qu'orienté méthodes.

Cela veut dire que dans de multiples situations, un résultat bon, rapide et facile à interpréter est plus utile qu'un processus parfait mais long et difficile. KXEN a développé de nombreuses techniques uniques pour automatiser la modélisation des données. La principale compétence nécessaire avant une utilisation de KXEN consiste à connaître les données que vous voulez analyser, et la problématique que vous souhaitez étudier : est-ce que votre problématique est de nature prédictive ou descriptive ? (ou dans un langage technique, est-ce une classification, une régression ou une problématique de Clustering ?)

## Pourquoi KXEN n'a pas d'arbres de décision, réseaux de neurones ou autres techniques ?

Dans les publications, vous rencontrerez une approche inverse. Cela comprend une plate-forme de Data Mining combinant une série de méthodes éprouvées (algorithmes) aussi nombreuses que possible. Leur message principal est le suivant : « Il n'y a pas une méthode unique pour résoudre de façon optimale tous les problèmes, donc vous devez les avoir toutes. »

Les équipes de KXEN sont en accord dans une certaine mesure avec cette idée. Avant de créer KXEN, ils ont aussi recherché le Saint Graal : l'algorithme parfait qui surpasserait tous les autres sur tous les types de données. A présent, ils ont évolué en

changeant leur perspective. Avec les récentes avancées mathématiques (notez : je n'ai pas dit statistiques), nous avons accompli une chose très importante :

C'est un processus automatisé de modélisation qui vous produit de BONS résultats PRESQUE A CHAQUE fois.

Ce processus automatisé recherché par KXEN est à comparer à un processus manuel de choix d'algorithmes qui avec les solutions « boîtes à outils » d'autres éditeurs demandent du temps (le temps de construire au moins un modèle avec chaque méthode) et beaucoup d'expertise (la maîtrise complète de tous les algorithmes est une compétence rare).

La voie menant à cette automatisation a été ouverte par des avancées remarquables en Mathématiques et Machine Learning. Vladimir Vapnik a été le fer de lance de ce mouvement avec sa Théorie sur l'Apprentissage Statistique. Il a été le premier à ouvrir la porte à de nouvelles manières de décomposer l'erreur faite dans les techniques de Machine Learning. Il a trouvé une structure dans l'expression de cette erreur et a abouti à des notions intéressantes qui permettent de structurer les techniques de modélisation. Qu'apporte cette structure spécifique ? En fait, au lieu de rechercher de façon non maîtrisée parmi toutes les méthodes du monde, elle vous donne une direction de recherche et compare les méthodes entre elles. Vous nous direz : « Donc vous avez toujours besoin de toutes les méthodes et de les comparer ? »

Non parce que le fait que nous ayons des fondations mathématiques et sachions ce que nous faisons, permet de dériver des meta-algorithmes qui effectueront cette recherche automatiquement. Toutes les étapes utilisées par KXEN Analytic Framework utilisent cette technique. Il y a des compromis à faire car cela doit être rapide et produire des résultats interprétables aisément.

L'architecture mathématique construit en interne plusieurs modèles qui concourent en même temps. Elle n'exécute pas de façon aléatoire un changement de techniques de modélisation ; elle utilise la Minimisation Structurale des Risques de Vapnik pour passer en revue des jeux de modèles. KXEN a développé une manière robuste de comparer les modèles avant de vous produire le modèle avec le meilleur compromis précision et robustesse (capacité de généralisation).

## **Les Algorithmes KXEN sont-ils propriétaires ?**

Comme le savent les experts, les algorithmes sont importants dans l'Analytique Prédicatif mais le problème clé est de rendre les données compatibles avec les algorithmes. Certains algorithmes n'acceptent que des symboles, d'autres que des nombres. Tous ceux qui ont de l'expérience, vous diront qu'ils passent beaucoup de temps sur la préparation des données et leur encodage.

En pratique, cela veut dire de nombreuses complications comme le traitement des valeurs manquantes, des valeurs hors normes, le meilleur encodage des données en fonction de l'algorithme choisi. Cela veut aussi dire avoir des algorithmes robustes, qui produisent régulièrement de bons résultats. C'est tout ?

Non, parce que vous souhaiteriez aussi que le traitement soit possible automatiquement sans mise au point manuelle de certains paramètres et que vous voudriez pouvoir extraire automatiquement de l'information de larges jeux de données avec beaucoup de colonnes. Alors comment KXEN a résolu cela ?

Premièrement grâce à son expertise dans le domaine de l'analyse de données provenant de la recherche et de l'expérience terrain. En second, KXEN a inclus des manières automatiques de traiter les valeurs manquantes et hors normes ainsi que l'encodage algorithmique. La préparation des données se fait en 2 passages. La première phase qui s'appelle aussi phase de manipulation de données permet aux experts de créer de nouvelles variables pertinentes pour le domaine. Par exemple, aucun système automatique ne vous dira que le dernier vendredi du mois est un bon indicateur pour les flux monétaires entre banques. La deuxième phase optimise l'encodage des attributs en fonctions des algorithmes utilisés.

L'objectif de KXEN est d'encoder automatiquement l'information et de façon optimale pour une problématique étudiée, une fois que l'expert du domaine l'a décrite.

Un autre facteur clé est l'aisance à interpréter les résultats. Tous les composants KXEN ont été conçus pour présenter des résultats pertinents à un utilisateur final et nous ne parlons pas d'un histogramme en 3D. Nous voulons dire les éléments qui seront affichés dans les graphiques comme la contribution des variables, l'importance des catégories, les indicateurs de qualité et fiabilité (robustesse) du modèle.

## Automatisation et Performance

Automatiser la construction de modèles finalement de piètre qualité aurait un intérêt considérablement réduit. Les modèles KXEN ont une très bonne capacité prédictive.

Les premiers clients de KXEN ont tous mené des benchmarks nombreux avant d'adopter nos solutions. La liste de clients prestigieux de KXEN constitue une première assurance de performances des modèles.

D'après notre expérience, la qualité (en terme de précision de la prévision) des modèles KXEN est parfaitement comparable avec celle des modèles construits avec du temps par des experts en statistiques sur des outils de type boîte à outils. Si dans 80% des cas, les performances prédictives sont identiques à 1 ou 2% près, il n'est pas rare que les modèles KXEN soient sensiblement meilleurs et ce à cause de deux propriétés fondamentales qui différencient les outils KXEN des outils classiques :

- Les algorithmes de KXEN permettent de construire des modèles de scores avec des milliers de variables en entrée sans altérer ni la robustesse (fiabilité) du modèle ni la lisibilité de son interprétation. Les algorithmes de statistiques classiques ne permettent pas d'avoir à la fois une lisibilité du modèle et l'utilisation de milliers de variables d'entrée. (On se reportera à l'article de Michel Bera ci-joint. Cette possibilité de traitement d'un grand nombre de variables d'entrée autorise les experts métier à construire de très nombreux indicateurs

- métier « potentiellement explicatifs » à partir des données disponibles et de laisser l'outil KXEN déterminer lesquels sont réellement explicatifs.
- Le codage des données est une étape longue et fastidieuse. Ainsi il est classique de voir des datamart d'analyse avec des codages de données réalisés une fois pour toutes. Ainsi le codage d'un code postal, difficile à réaliser, sera fait une fois pour toutes en regroupant par exemple les codes postaux en département. Le codage automatique réalisé par KXEN prendra alors toute sa force, car le codage sera toujours fait en fonction de la problématique traitée. Et le regroupement pourra montrer par exemple une dissimilarité entre codes postaux des régions rurales, urbaines ou péri-urbain, regroupement qui n'épouse absolument pas un découpage départemental.

Enfin et ce sera notre dernier point, la communauté des chercheurs et des praticiens en théorie statistique de l'apprentissage est en essor constant depuis une dizaine d'années. Ainsi cette théorie est aujourd'hui enseignée dans de nombreuses écoles et universités dans le Monde. Cette théorie est enseignée à l'ENSAE, au CNAM, au MIT, à Stanford ....

KXEN a constitué un comité scientifique dirigé par Michel Béra, co-fondateur de KXEN, normalien, docteur en statistique non paramétrique et ancien assistant du professeur Benzécri. Ce comité qui assiste KXEN dans ses travaux de recherche et développement regroupe aujourd'hui des experts mondialement reconnus en data mining et statistique :

Gilbert Saporta (titulaire de la chaire de Statistiques appliquées au CNAM),  
Bernhard Schölkopf (Max Planck Institute en Allemagne)  
Grégory Shapiro (Fondateur de KDnuggets.com aux USA),  
Emmanuel Viennet (Maître de Conférences à l'Université Paris 13),  
Léon Bottou (NEC Research Institute à Princeton),  
Yann LeCun (NEC Research Institute à Princeton),