**UNIVERSITE DE LISALA**

**CENTRE INTERUNIVERSITAIRE DE RECHERCHE PLURIDISCIPLINAIRE (CIREP)**
**STATUT : UNIVERSITE PUBLIQUE**
**Web : www.cirep.ac.cd**
**Email : info@cirep.ac.cd**

# NOTES DE COURS DE DATA MINING

## OBJECTIFS DU COURS

### *Objectif général*

L'objectif général de ce cours est d'enseigner aux étudiants les principes, les techniques et les applications de l'exploration de données pour extraire des informations utiles à partir de grands ensembles de données. Les objectifs spécifiques peuvent varier en fonction du niveau du cours et du public cible.

### *Objectifs spécifiques :*

- ✓ Comprendre les concepts fondamentaux du data mining, y compris les types de données, les techniques d'exploration et les applications pratiques.
- ✓ Maîtriser les différentes étapes du processus de data mining, de la collecte des données à l'interprétation des résultats.
- ✓ Acquérir des compétences pratiques dans l'utilisation d'outils et de logiciels de data mining pour analyser des ensembles de données réels.
- ✓ Apprendre à évaluer la qualité des modèles de data mining et à choisir la meilleure approche en fonction des objectifs de l'analyse.
- ✓ Explorer les principaux algorithmes de data mining pour la classification, la prédiction, le clustering et l'association de données.
- ✓ Développer des compétences en interprétation des résultats du data mining et en communication des conclusions aux parties prenantes.
- ✓ Examiner les défis éthiques et les implications sociales du data mining, notamment en ce qui concerne la confidentialité des données et la protection de la vie privée.
- ✓ Explorer les applications pratiques du data mining dans divers domaines, tels que le marketing, la finance, la santé, la recherche scientifique et d'autres secteurs.

# I) Introduction

Qu'est-ce que le data Mining ?

Extraction d'informations intéressantes (non triviales, implicites, préalablement inconnues et potentiellement utiles) à partir de grandes bases de données. C'est analyser les données pour trouver des patrons cachés en utilisant des moyens automatiques.

C'est un processus non élémentaire de recherche de relations, corrélations, dépendances, associations, modèles, structures, tendances, classes (clusters), segments, lesquelles sont obtenues de grande quantité de données (généralement stockées sur des bases de données (relationnelles ou no)). Cette recherche est effectuée à l'aide des méthodes mathématiques, statistiques ou algorithmiques.

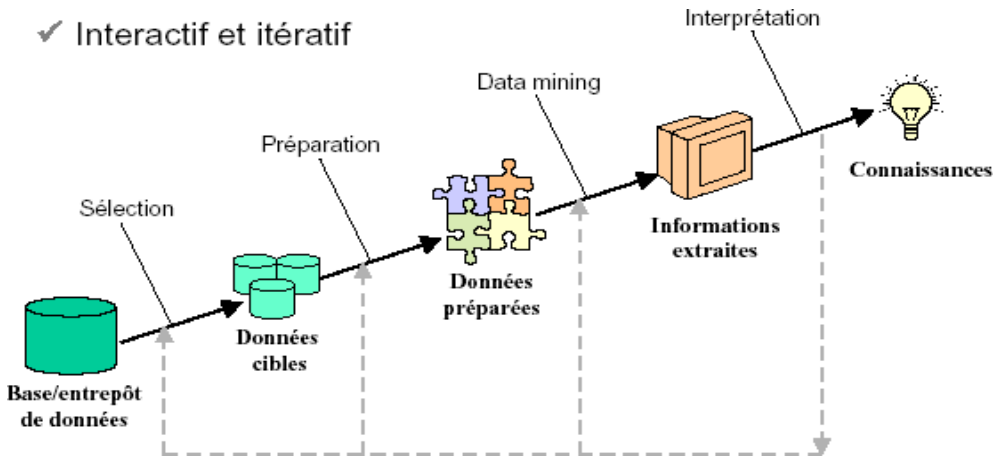
Data Mining se considère comme un processus le plus automatique possible, qui part de données élémentaires disponibles dans un Data Warehouse à la décision. L'objectif principale de Dat Mining c'est de créer un processus automatique qui a comme point de départ les données y comme finalité l'aide à la prise des décisions.

Data Mining versus KDD (Knowledge Discovery in Databses)

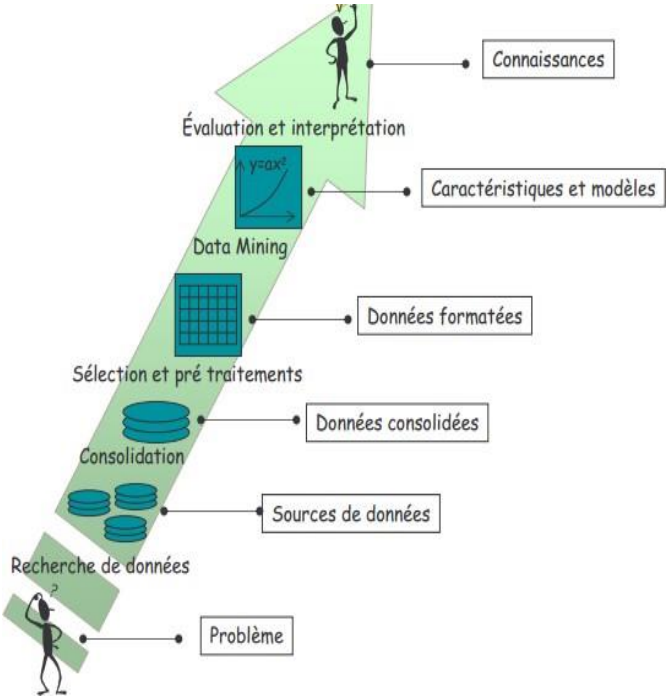
- habituellement les deux termes sont interchangés.
- KDD (Knowledge Discovery in Databses) : C'est le processus de trouver information et/ou partons utiles à partir de données.
- Data Mining : C'est l'utilisation des algorithmes pour extraire information et/ou partons comme partie du processus KDD.

Data Mining: C'est une partie du processus KDD

Data Mining: Le cœur du processus d'extraction de connaissances.




Processus KDD




<u>Statistique vs Data mining</u>

- En statistique :
  - Quelques centaines d'individus
  - Quelques variables
  - Fortes hypothèses sur les lois statistiques
  - Importance accordée au calcul
  - Échantillon aléatoire.
- En Data mining
  - Des millions d'individus
  - Des centaines de variables
  - Données recueillies sans étude préalable
  - Nécessité de calculs rapides

- Corpus d'apprentissage.

Data Mining versus Data Warehouse

Data warehouse est un entrepôt de données d'une entreprise qui contient quelques données opérationnelles, données agrégées (agrégations), données historiques, données évolutives et possiblement des données externe à l'entreprise qui ont une relation avec l'activité de l'entreprise. Ces données sont stockées dans une ou plusieurs base de données relationnelle et sont accessibles par toutes les applications orientées aide à la décision.

Évidemment Data Warehouse et Data Mining sont deux choses très différentes. Data Warehouse est usuellement le point le départ de Data Mining. Data Warehouse et Data Mining sont des parties du processus KDD.

Qu'est-ce que le Data Warehouse

| Caractéristiques | BDD | DataWarehouse |
|---|---|---|
| Utilisation | SGBD (base de production) | Datawarehouse |
| Opération typique | Mise à jour | Analyse |
| Type d'accès | Lecture écriture | Lecture |
| Niveau d'analyse | Elémentaire | Global |
| Quantité d'information échangées | Faible | Importante |
| Orientation | Ligne | Multidimension |
| Taille BD | Faible (max qq GB) | Importante (pouvant aller à plusieurs TB). |
| Ancienneté des données | Récente | Historique |

Data Mining versus Machine Learning

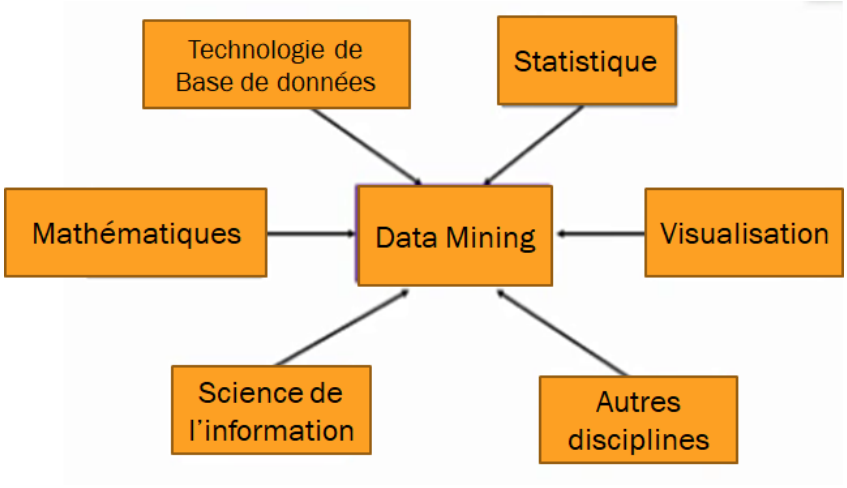
Machine Learning: C'est un sujet de l'intelligence artificielle (IA) qui s'occupe de la façon d'écrire des programmes qui peuvent apprendre. Dans Data Mining machine learning est habituellement utilisés pour la prédiction et classification.Machine learning se divise en deux : Apprentissage supervisé (learn by example) et apprentissage non supervisé.

Data Mining: sur quels types de données

- Fichiers plats
- BD's relationnelles
- Data warehouses

- BD's transactionnelles
- BD's avancées
  - BD's objet et objet-relationnelles
  - BD's spatiales
  - Séries temporelles
  - BD's Textes et multimedia
  - BD's Hétérogènes
  - WWW (web mining)

Data Mining: Intersection de multiples disciplines




Applications par domaine

| | |
|---|---|
| Services financiers<br>– Attrition (churn)<br>– Détection de fraudes<br>– Identification opportunités de ventes | Marketing<br>– Gestion de la relation client (CRM)<br>– Optimisation de campagnes marketing<br>– Ventes croisées |
| Télécommunications<br>– Fidélisation (anti-churn)<br>– Ventes croisées<br>– Incidentologie | Assurances, Secteur public<br>– Indiquer les anomalies des comptes<br>– Réduire le coût d'investissement d'activité suspecte<br>– Détection de la fraudes |
| Grande Distribution<br>– Fidélisation<br>– Ventes croisées<br>– Analyses de panier<br>– Détection de fraudes | Sciences de la vie<br>– Trouver les facteurs de diagnostic typiques d'une maladie<br>– Alignement gênes & protéïnes<br>– Identifier les capacités d'interaction de médicaments |
| Internet<br>– Personnalisation des pub affichées<br>– Optimisation des sites web<br>– Profilage et Recommendation | Autre<br>– Rech. d'info (web ou document)<br>– Recherche par similarité (images…)<br>– Analyse spatiale… |

**Comparaison Marketing traditionnel – Marketing one-to-one**

| Marketing traditionnel | Marketing one-to-one |
|---|---|
| Client anonyme | Client individualisé |
| Produit standard | Produit et service personnalisés |
| Production en série | Production sur mesure |
| Publicité à large diffusion | Message individuel |
| Communication unilatérale | Communication interactive |
| Réalisation d'une vente | Fidélisation du client |
| Part de marché | Part de client |
| Large cible | Niche rentable |
| Canaux de distribution traditionnels | Nouveaux canaux (plates-formes téléphoniques, Internet, téléphones mobiles) |
| Marketing orienté « produit » | Marketing orienté « client » |

Pourquoi utiliser Data Mining ?

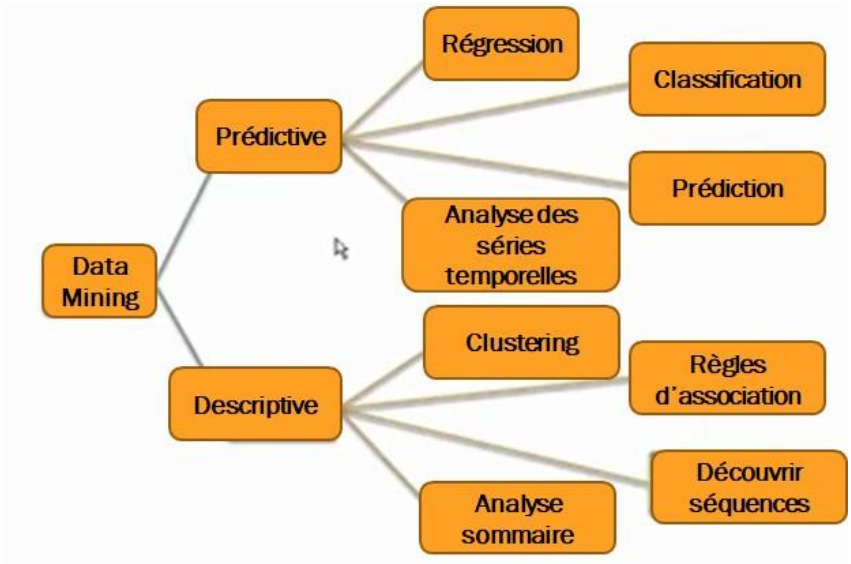
- Problème de l'explosion de données

Les outils automatiques de collecte de données font que les Bases de Données (BD's) contiennent énormément de données (Ex : La base de données des transactions d'un super marché).

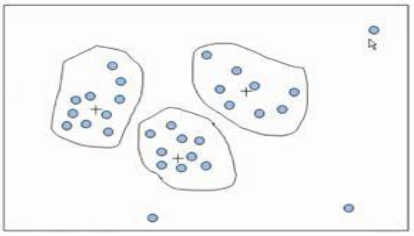
- Beaucoup de données mais peu de connaissances !
- Solution: Data warehousing et data mining
    - Data warehousing et OLAP (On Line Analytical Processing)
    - Extraction de connaissances intéressantes (règles, régularités, patterns, contraintes) à partir de données

Tâches réalisées en Data Mining

- Descriptives : consiste à trouver les caractéristiques générales relatives aux données fouillées (Résumé/synthèse, Clustering, Règles d'association)
- Prédictives : Consiste à utiliser certaines variables pour prédire les valeurs futures inconnues de la même variable ou d'autres variables (Séries temporelles, Régression, Classification).




- **Clustering :** (classification non supervisée, apprentissage non supervisé) : c'est similaire à la classification, sauf que les groupes no sont pas prédéfinies. L'objectif est de décomposer ou de segmenter un ensemble de données ou individus en groupes qui peuvent être disjoints ou non.
  - Les groupes se forment à base de la similarité des données o des individus en certaines variables.
    - Comme groupes suggérés (imposés) par les données, pas définis a priori l'expert doit donner une interprétation des groupes qui se forment.



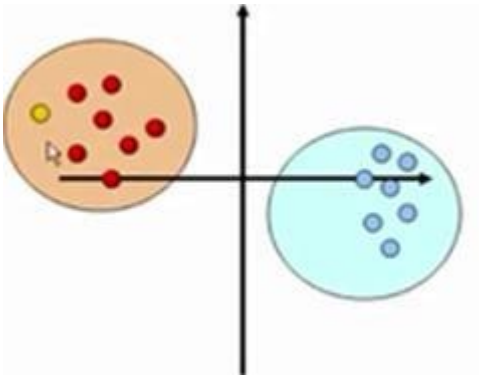
- Méthodes :
  - K-means
  - Classification hiérarchique (groupes disjoints)
  - nuées dynamiques (groupes disjoints)
  - Classification pyramidale (groupes non disjoints)

- **Classification :**(discrimination) : associer des données à des groupes prédéfinis (apprentissage supervisé)

Trouver des modèles (fonctions) qui décrivent et distinguent des concepts pour de futures prédictions.

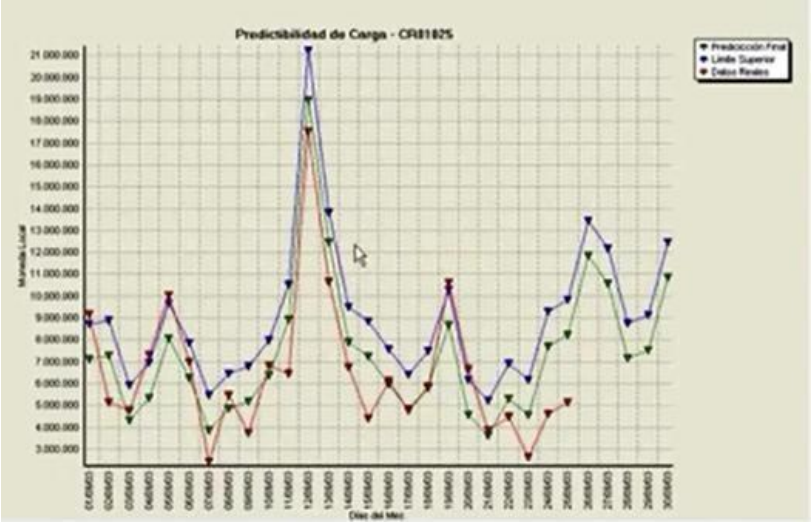
- Méthodes :
  - Arbres de décision,

- règles de classification,
- réseaux neuronaux.



**Régression :**la régression est utilisée pour prédire les valeurs absentes d'une variable en se basant sur sa relation avec les autres variables de l'ensemble de données.

- Régression linéaire, non linéaire, logistique, logarithmique, univariée, multivariée, entre d'autres.




**Règles d'association (analyse d'affinité) :** connue comme (Link Analysis) se réfère à découvrir les relations non évidentes entre les données.

- Méthodes :
  - Règles d'associations (association rules)
  - Analyse de corrélation et de causalité

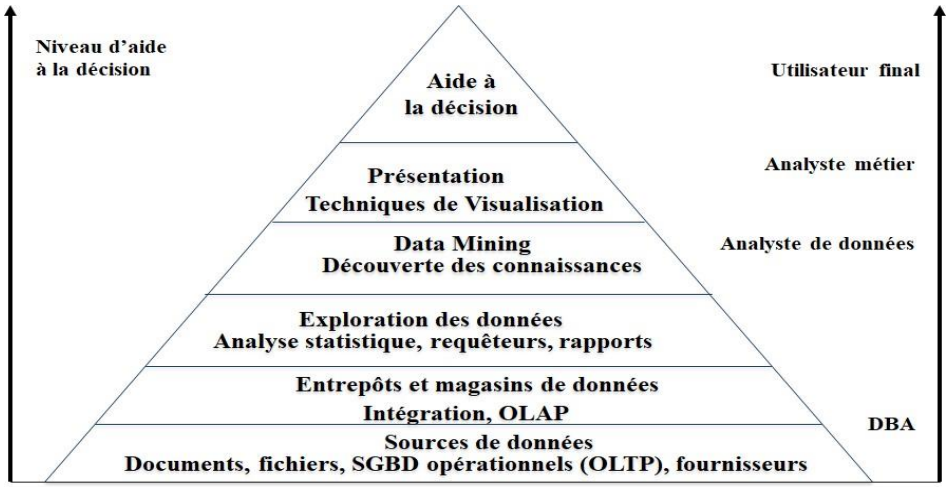
Business Intelligence

Business Intelligence (BI) est un concept proposé par IBM, Microsoft, Oracle, … pour *« Consolider la quantité gigantesque de données atomiques que les entreprises génèrent en information pour que les gens puissent les accéder, les comprendre et les utiliser »*

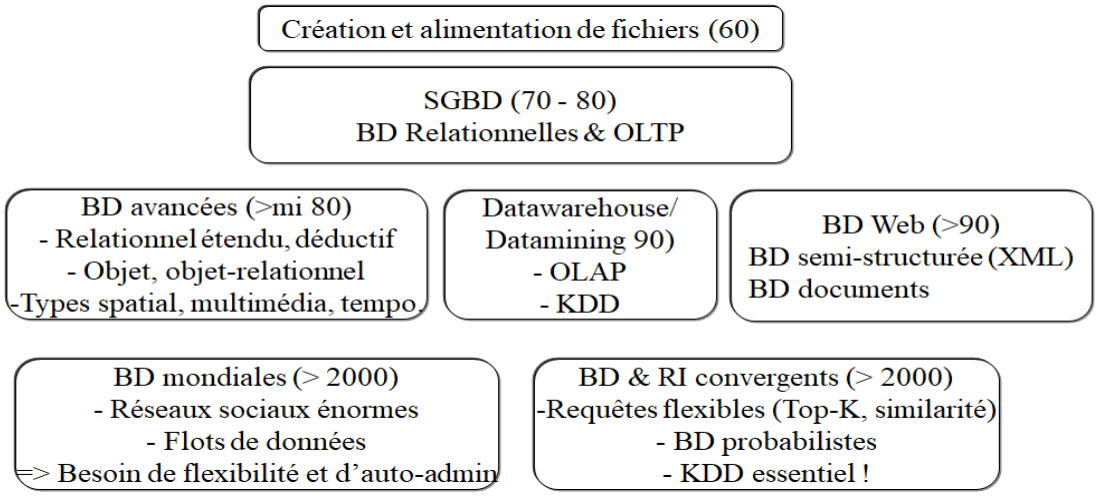
*=> Présenter l'information dans des formats plus utiles, en utilisant des outils d'exploration, de reporting et de visualisation avancés.*

Son but est d'améliorer les performances décisionnelles de l'entreprise en répondant aux demandes d'analyse des décideurs *non informaticiens et non statisticiens*

Pyramide de la BI




Historique




- Le data mining n'est pas nouveau :
  - 1875 : Régression linéaire
  - 1936 : Analyse discriminante
  - 1943 : Réseaux de neurone
  - 1944 : Régression logistique
  - 1984 : Arbres de décision
  - 1990 : Apparition du concept de data mining

Cycle de vie d'un projet de Data Mining

1. Apprentissage du domaine d'application :
    - Connaissances nécessaires et buts de l'application
2. Création du jeu de données cible : sélection des données
3. Nettoyage et prétraitement des données (jusqu'à 60% du travail !)
4. Réduction et transformation des données
    - Trouver les caractéristiques utiles, dimensionnalité/réduction des variables
5. Choix des fonctionnalités data mining
    - synthèse, classification, régression, association, clustering
6. Choix des algorithmes
7. Data mining : recherche de motifs (patterns) intéressants
8. Évaluation des motifs et représentation des connaissances
    - visualisation, transformation, élimination des motifs redondants, etc.
9. Utilisation des connaissances découvertes.

Ce qui n'est pas de Data Mining

- En générale Data Mining n'est pas basé sur des modèles déterministes.
- Un modèle déterministe ne fait intervenir aucune variable aléatoire. Les relations entre variables sont strictement fonctionnelles.

Ce qui n'est pas de la fouille de données

- En générale Data Mining est basé sur des modèles **probabilistes**.
- Un modèle probabiliste est un modèle mathématique qui nous aide à prévoir le comportement des futures répétitions d'une expérience aléatoire en se basant sur l'estimation d'une probabilité d'apparition de cet évènement concret.

# Chapitre 1
# Histoire et installation de R

# II) Histoire et installation de R

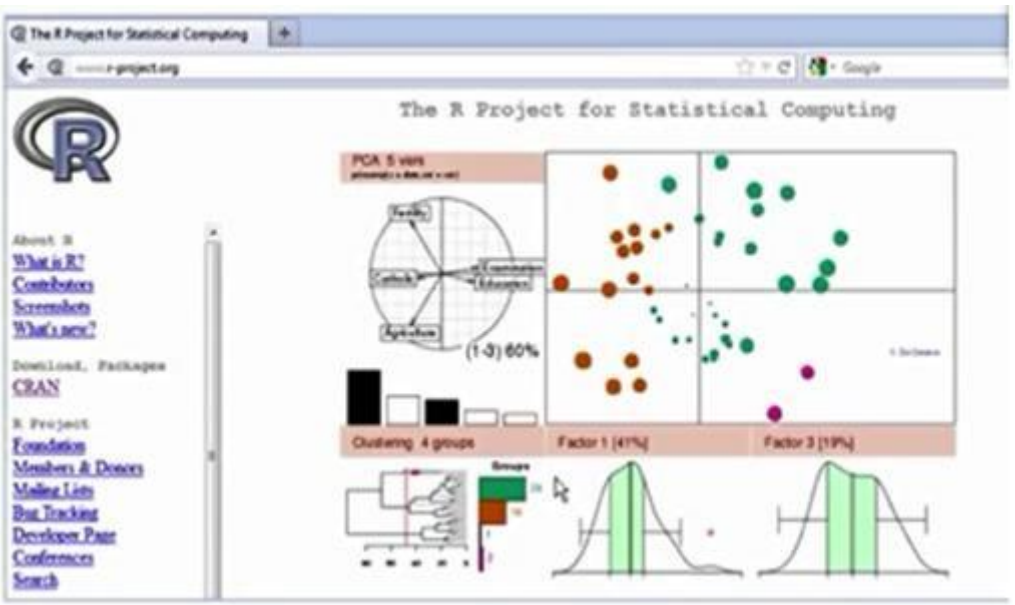
***R*** est un clône **gratuit** du logiciel **S-Plus** commercialisé par MathSoft, développé par Statistical Sciences autour du langage S (conçu par les laboratoires Bell).

***S*** a été crée par le professeur

**John M. Chambers** et son équipe de l'Université de Stanford.

R a été créé par Ross Ihaka et Robert Gentleman à l'Université d'Auckland, Nouvelle Zélande, et est maintenant développé par la R Development Core Team.

## II.1 - R Project

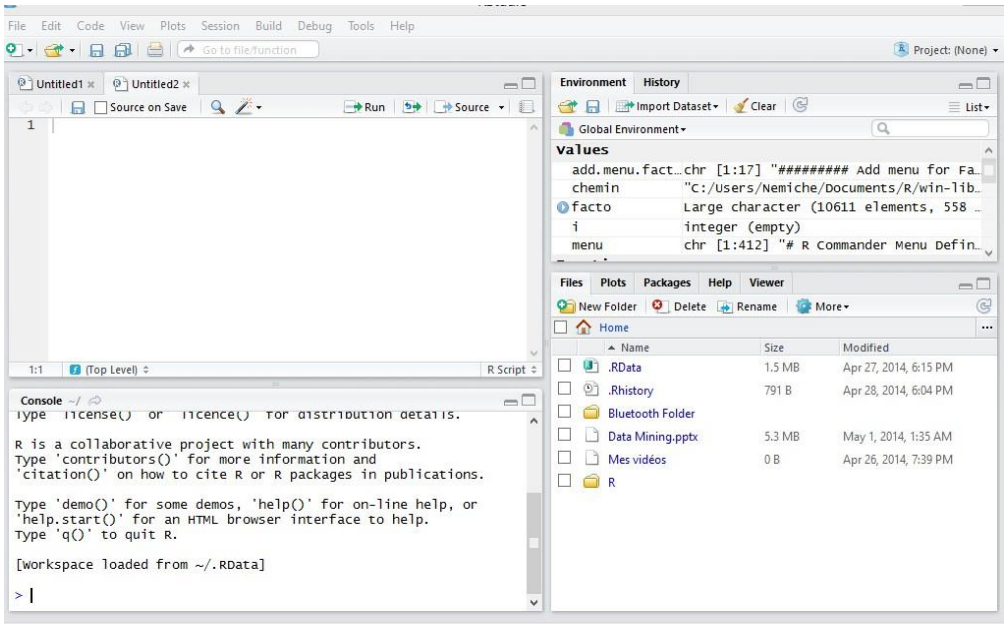



Installation de R

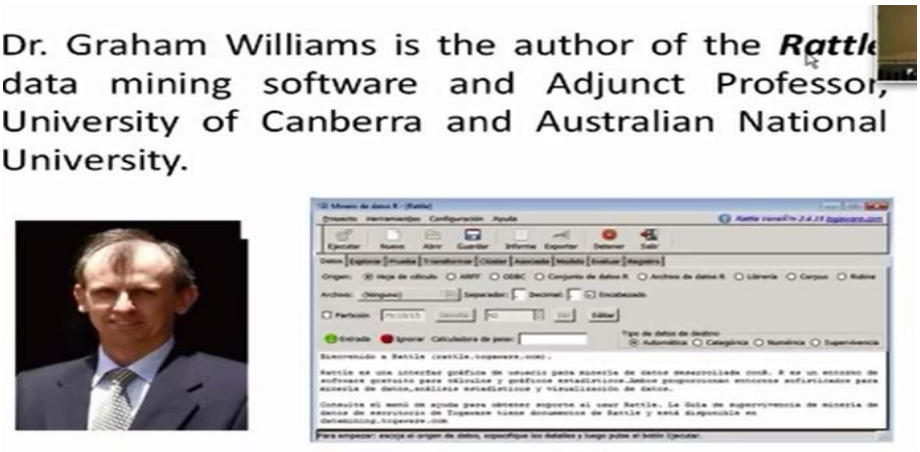
1. Rendez-vous sur le site http://www.r-project.org/
2. Puis, à gauche sur la page d'accueil, vous trouverez un menu Download, Packages. Dans ce menu, cliquez sur CRAN.

## II.2 - Rstudio

http://www.rstudio.com

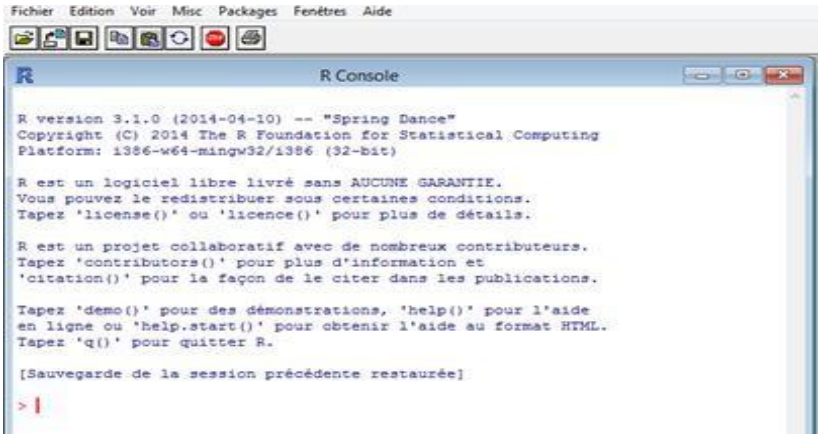



## II.3 - Rattle

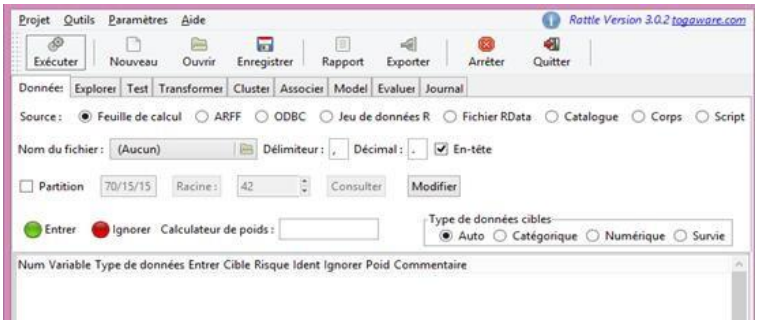



- Pour l'installer
  - install.packages(« rattle »,dependencies=TRUE)
- Pour l'éxecuter :
  - library (rattle)
  - rattle ()
- Site web :
  - http://rattle.togaware.com/

Interface de R sous Windows

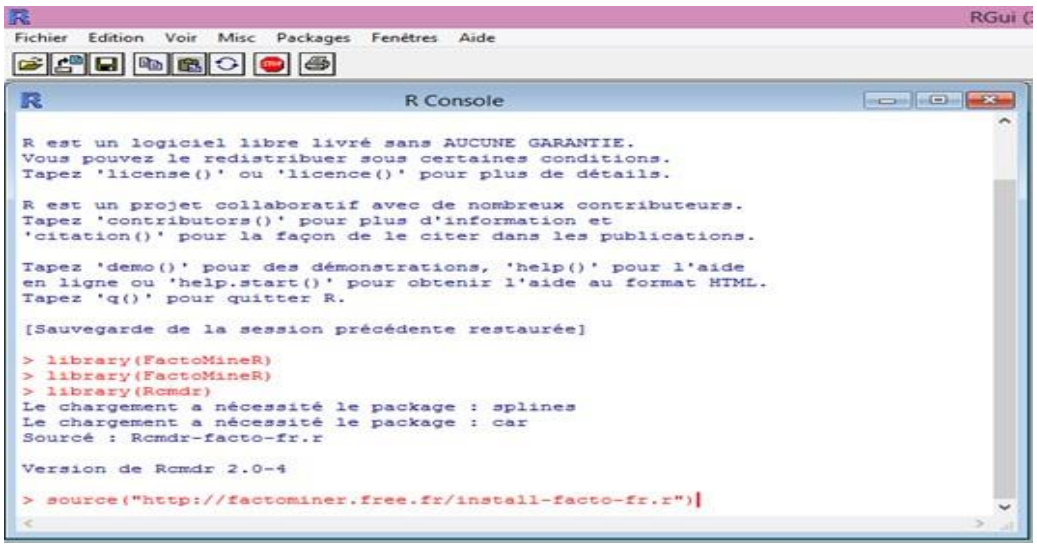



Interface de Rattle

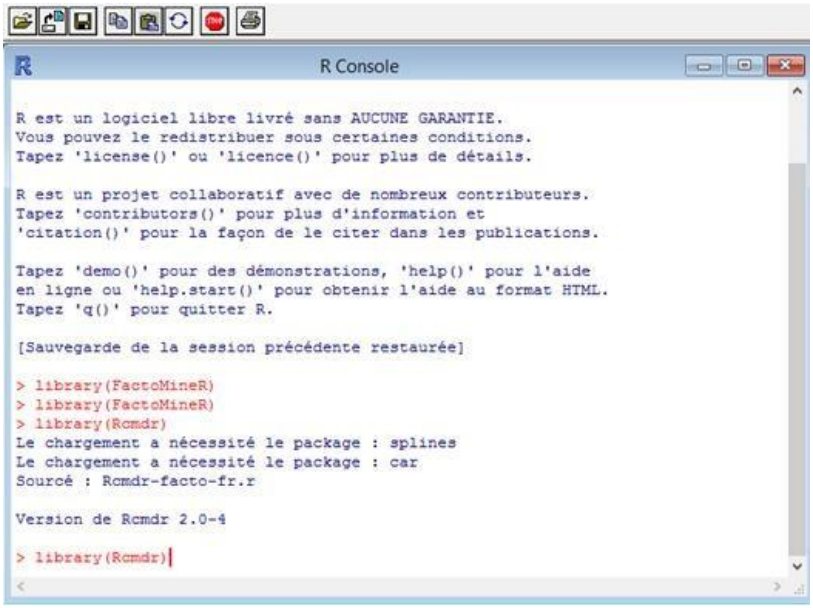


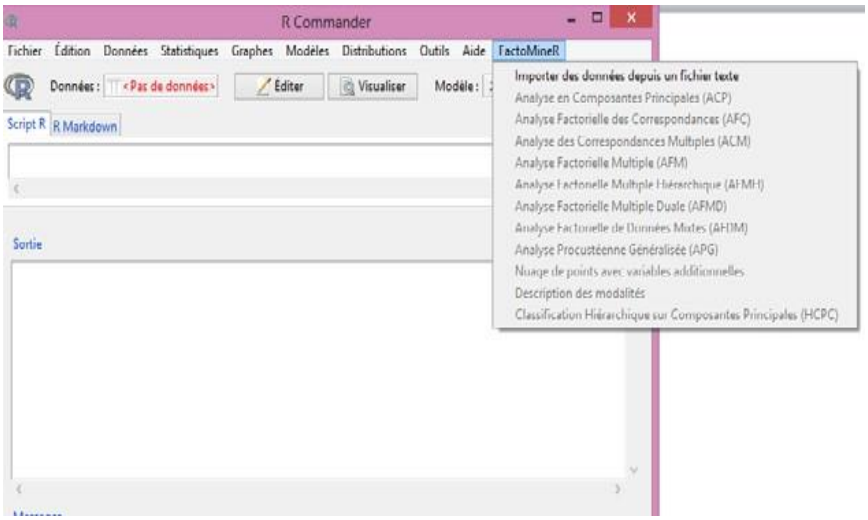

## II.4 - FactoMineR

- FactoMineR a été créé dans le département de Mathématiques Appliquées de: Agrocampus de l'Université de Rennes, France.
- Vous avez la possibilité d'installer **FactoMineR** comme un package classique ou d'installer **FactoMineR** et son interface graphique afin de l'utiliser de façon plus conviviale:http://factominer.free.fr/index_fr.html
- Pou installer FacoMineR GUI :source("http://factominer.free.fr/install-facto-fr.r")



Utilisationde FactoMineR sur Rcommander




Introduction à Rcommander

- Rcommander a été créé par John Fox et son équipe, c'est une interface graphique qui couvre la plupart d'analyses statistiques habituelles.

C'est une manière d'utiliser R sans nécessité d'apprendre le code (utiliser).

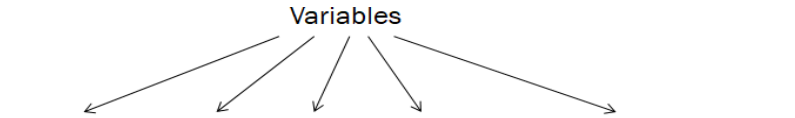
# Chapitre 2
# Analyse exploratoire

# III) Analyse exploratoire (descriptive)

Une variable est une propriété ou caractéristique d'un individu

- Exemple : Couleur des yeux d'une personne, température, état civil, …
- Une collection de variables décrivant à un individu

On dit individu ou enregistrement, point, cas, objet, entité, exemple d'observation




| age | Revenus | Etudiant | Taux_crédit | Achat_PC |
|---|---|---|---|---|
| <=30 | élevé | non | faible | non |
| <=30 | élevé | non | excellent | non |
| 31…40 | élevé | non | faible | oui |
| >40 | moyen | non | faible | oui |
| >40 | faible | oui | faible | oui |
| >40 | faible | oui | excellent | non |
| 31…40 | faible | oui | excellent | oui |
| <=30 | moyen | non | faible | non |
| <=30 | faible | oui | faible | oui |
| >40 | moyen | oui | faible | oui |
| <=30 | moyen | oui | excellent | oui |
| 31…40 | moyen | non | excellent | oui |
| 31…40 | élevé | oui | faible | oui |
| >40 | moyen | non | excellent | non |

## III.1 - Types de variables

Qualitative : les variables représentent des catégories différentes au lieu des numéros. Les opérations mathématiques comme la somme et la soustraction n'ont pas de sens.

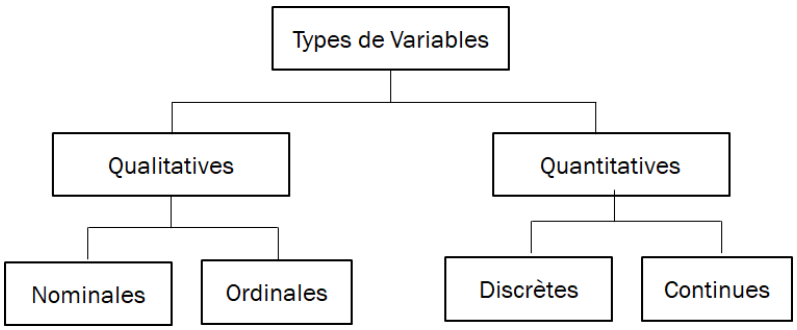
- Exemples : couleur des yeux, niveau académique, adresse IP

Quantitative : les variables sont les numéros

- Exemple : poids, la température, le nombre d'enfants

### III.1.1) Variables qualitatives

| ind | SEXO | EDAD | INGRESO |
|---|---|---|---|
| 1 | F | 5 | Medio |
| 2 | F | 3 | Alto |
| 3 | M | 4 | Bajo |
| 4 | F | 1 | Bajo |
| 5 | M | 2 | Medio |
| 6 | M | 5 | Alto |
| 7 | F | 2 | Medio |
| 8 | M | 3 | Bajo |
| 9 | M | 1 | Alto |
| 10 | F | 4 | Medio |




### III.1.2) Transformation d'une variable quantitative envariable qualitative

Pour les variables discrètes : considérer que les valeurs prises par la variable sont les modalités de la variable qualitative (ordonnée)

- Pour les variables continues :
  - on divise l'intervalle [a ; b[ où varie la variable en un certain nombre d'intervalles [a ; x1[, [x1 ; x2[, [xi; x i+1[ … , [x p-1; b[ et
  - on dénombre pour chaque intervalle le nombre d'individus dont la mesure appartient à l'intervalle
  - En règle générale, on choisit des classes de même amplitude.
  - Pour que la distribution en fréquence soit intéressante, il faut que chaque classe comprenne un nombre « suffisant » d'individus (ni)
  - Si la longueur des intervalles est trop grande, on perd trop d'information

Il existe des formules empiriques pour établir le nombre de classes pour un échantillon de taille n

- Règle de Sturge
  - Nombre de classes =1 + 3.3 log n
- Règle de Yule

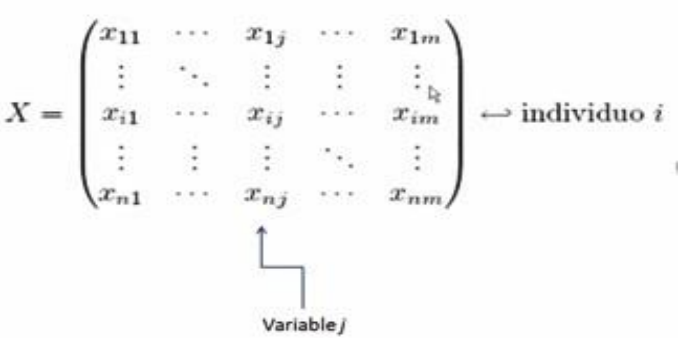
- Nombre de classes = 2.5√ n

- L'intervalle entre chaque classe est calculé par
  - (b-a)/nombre de classes
- On calcule ensuite à partir de a les classes successives par addition.

NB: il n'est pas obligatoire d'avoir des classes de même amplitude. Mais pas de chevauchement d'intervalle

## III.2 - Les données

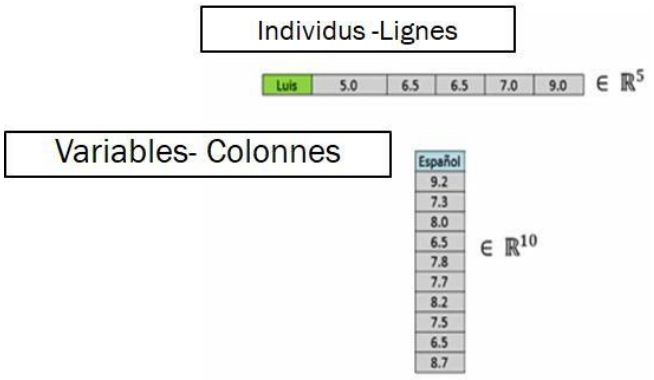
- Le point de départ est d'une table de données:



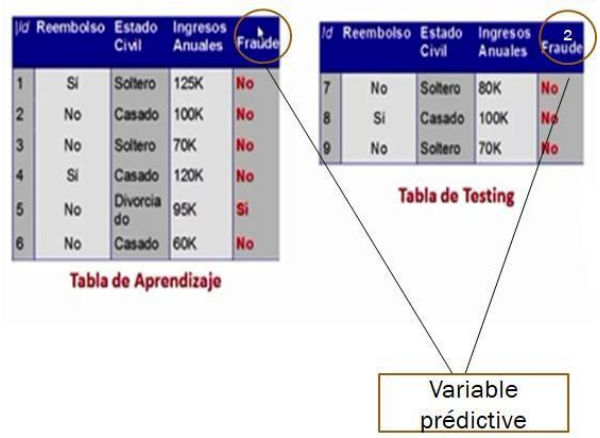

Exemple

| | Matemáticas | Ciencias | Español | Historia | EdFísica |
|---|---|---|---|---|---|
| Lucía | 7.0 | 6.5 | 9.2 | 8.6 | 8.0 |
| Pedro | 7.5 | 9.4 | 7.3 | 7.0 | 7.0 |
| Inés | 7.6 | 9.2 | 8.0 | 8.0 | 7.5 |
| Luis | 5.0 | 6.5 | 6.5 | 7.0 | 9.0 |
| Andrés | 6.0 | 6.0 | 7.8 | 8.9 | 7.3 |
| Ana | 7.8 | 9.6 | 7.7 | 8.0 | 6.5 |
| Carlos | 6.3 | 6.4 | 8.2 | 9.0 | 7.2 |
| José | 7.9 | 9.7 | 7.5 | 8.0 | 6.0 |
| Sonia | 6.0 | 6.0 | 6.5 | 5.5 | 8.7 |
| María | 6.8 | 7.2 | 8.7 | 9.0 | 7.0 |

## III.3 - Nuage de points



Données pour les méthodes prédictives




Exemple

| | Matemáticas | Ciencias | Español | Historia | EdFísica | Tipo |
|---|---|---|---|---|---|---|
| Lucía | 7.0 | 6.5 | 9.2 | 8.6 | 8.0 | Regular |
| Pedro | 7.5 | 9.4 | 7.3 | 7.0 | 7.0 | Bueno |
| Inés | 7.6 | 9.2 | 8.0 | 8.0 | 7.5 | Bueno |
| Luis | 5.0 | 6.5 | 6.5 | 7.0 | 9.0 | Malo |
| Andrés | 6.0 | 6.0 | 7.8 | 8.9 | 7.3 | Regular |
| Ana | 7.8 | 9.6 | 7.7 | 8.0 | 6.5 | Bueno |
| Carlos | 6.3 | 6.4 | 8.2 | 9.0 | 7.2 | Regular |
| José | 7.9 | 9.7 | 7.5 | 8.0 | 6.0 | Bueno |
| Sonia | 6.0 | 6.0 | 6.5 | 5.5 | 8.7 | Regular |
| María | 6.8 | 7.2 | 8.7 | 9.0 | 7.0 | Malo |

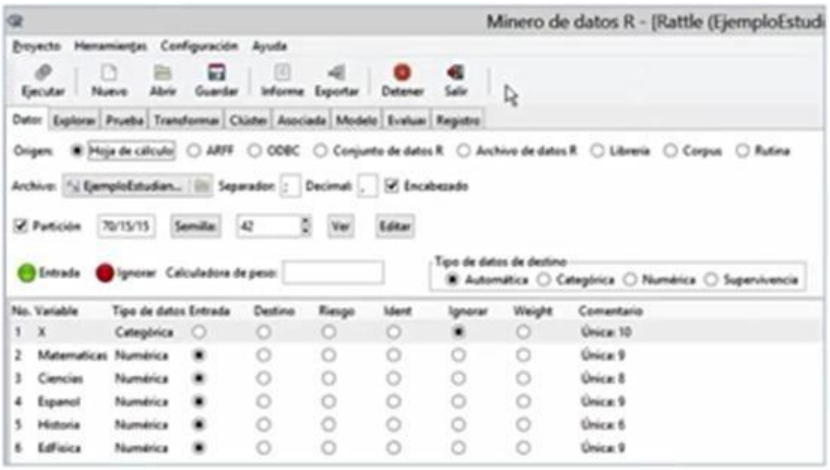


Comment lire des données en R?

| | Matemáticas | Ciencias | Español | Historia | EdFísica |
|---|---|---|---|---|---|
| Lucía | 7.0 | 6.5 | 9.2 | 8.6 | 8.0 |
| Pedro | 7.5 | 9.4 | 7.3 | 7.0 | 7.0 |
| Inés | 7.6 | 9.2 | 8.0 | 8.0 | 7.5 |
| Luis | 5.0 | 6.5 | 6.5 | 7.0 | 9.0 |
| Andrés | 6.0 | 6.0 | 7.8 | 8.9 | 7.3 |
| Ana | 7.8 | 9.6 | 7.7 | 8.0 | 6.5 |
| Carlos | 6.3 | 6.4 | 8.2 | 9.0 | 7.2 |
| José | 7.9 | 9.7 | 7.5 | 8.0 | 6.0 |
| Sonia | 6.0 | 6.0 | 6.5 | 5.5 | 8.7 |
| María | 6.8 | 7.2 | 8.7 | 9.0 | 7.0 |

Fichier texte CSV


Chargement de données en Rattle




## III.4 - Description d'une variable quantitative

Une variable quantitative est décrite par les valeurs qui prennent l'ensemble de n individus pour lesquels a été définis

Exemple

| individuo | tamaño |
|---|---|
| 1 | 1.70 |
| 2 | 1.65 |
| 3 | 1.70 |
| 4 | 1.80 |

Pour résumer l'information d'une variable quantitative les indices les plus communes sont :

- La moyenne. Définit par :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- La Variance : définit par

$$\mathrm{var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2$$

- L'écart type :

$$\sigma_x = \sqrt{\mathrm{var}(X)}$$

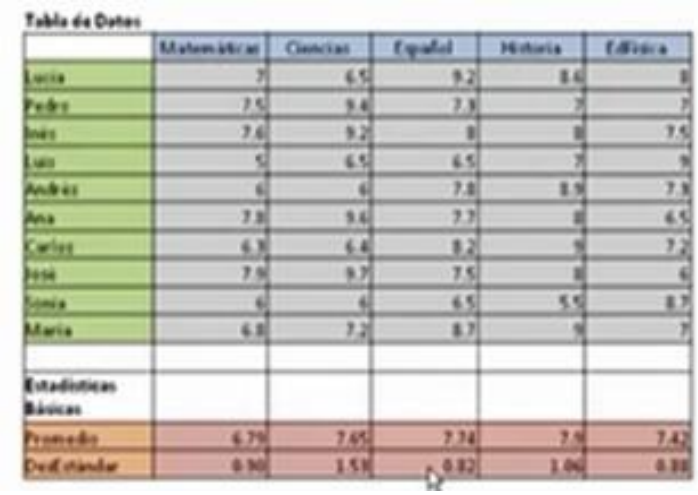


Tabla de Datos

| | Matemáticas | Ciencias | Español | Historia | Física |
|---|---|---|---|---|---|
| Lucia | 7 | 6.5 | 9.2 | 8.6 | 8 |
| Pedro | 7.5 | 9.4 | 7.3 | 7 | 7 |
| Inés | 7.6 | 9.2 | 8 | 8 | 7.5 |
| Luis | 5 | 6.5 | 6.5 | 7 | 9 |
| Andres | 6 | 6 | 7.8 | 8.9 | 7.3 |
| Ana | 7.8 | 9.6 | 7.7 | 8 | 6.5 |
| Carlos | 6.3 | 6.4 | 8.2 | 9 | 7.2 |
| José | 7.9 | 9.7 | 7.5 | 8 | 6 |
| Sonia | 6 | 6 | 6.5 | 5.5 | 8.7 |
| Maria | 6.8 | 7.2 | 8.7 | 9 | 7 |
| | | | | | |
| Estadísticas Básicas | | | | | |
| Promedio | 6.79 | 7.65 | 7.74 | 7.9 | 7.42 |
| DesEstandar | 0.90 | 1.53 | 0.82 | 1.06 | 0.88 |

- Le Coefficient de détermination :

**$R^2$ = Var (estimés par l'équation de régression)/Var (totale)**

$$R^2 = \frac{var(aX+b)}{var(Y)}$$

- Le Coefficient de corrélation :

$$R = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

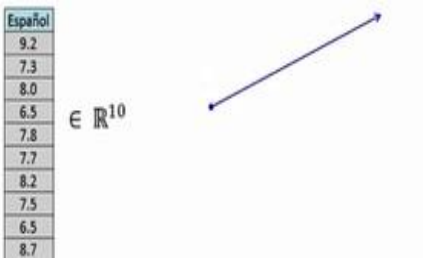
## III.5 - Matrice de Corrélation

Grande corrélation positive implique que si une variable augmente l'autre aussi augmente. Grande corrélation négative implique que si une variable augmente l'autre diminue et vice versa.

Corrélation proche de 0 implique l'absence de relation entre les variables

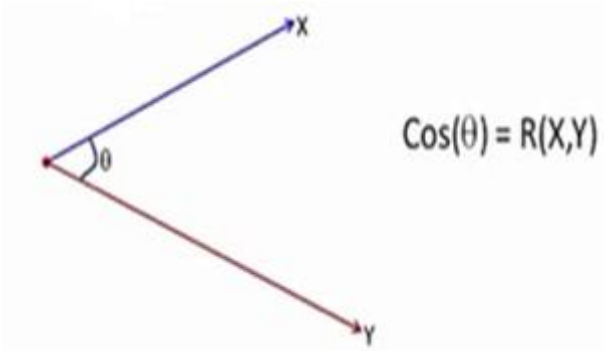
### III.5.1) <u>Interprétation géométrique du coefficient de corrélation</u>

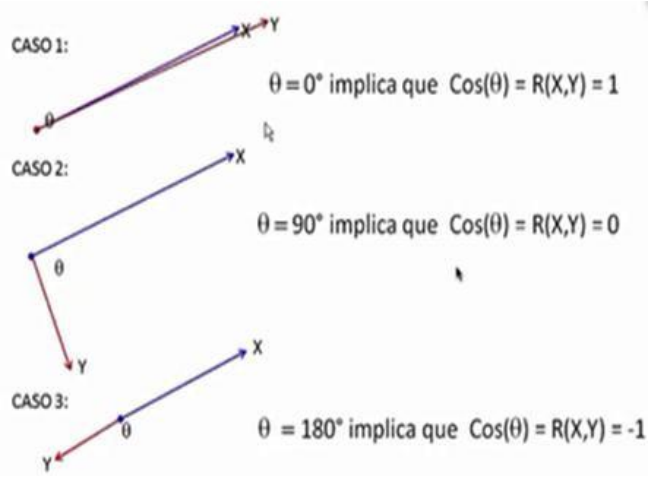
- Une variable x qui prend n valeurs peut être représenter comme un vecteur de $R^n$
- Variables -colonnes




- Théorème :

Dans l'espace vectoriel des variables $R^n$ [le] cosinus de l'angle entre 2 variables réduites er centrées est égale au coefficient de corrélation entres ses deux variables :



# Chapitre 3
# Règles d'Association

## IV) Les Règles d'Association Concepts Basiques

- En data mining, on utilise la technique des règles d'association pour déterminer les éléments qui se retrouvent ensembles.
- L'analyse du panier d'épicerie (« market basket analysis ») est un terme plus spécifique au commerce au détail. Cette analyse utilise les règles d'association.
- Dans une épicerie, les règles d'association décrivent les produits qui se retrouvent dans le même panier.

**Beurre d'arachides** ⇨ **Pain en tranches**

- Définitions
  - **Transactions** : achats fait par un seul client.
  - **Items** : produits achetés.
  - **Règle d'association** : énoncé de la forme (item X) ⇨ (item Y).
    - Item X = produit à analyser
    - Item Y = produit associé

- Règle d'association :
  - On choisira d'étudier des règles d'association permettant d'en apprendre davantage sur le comportement des clients. Les résultats de l'analyse devront être utiles et pratiques.
  - On choisira un niveau de granularité. On peut étudier l'association entre des ensembles de produits : ceux qui achètent des céréales achètent aussi du lait.

- La force d'association sera mesurée par :
  - **Support** : probabilité d'acheter le produit X et le produit Y.

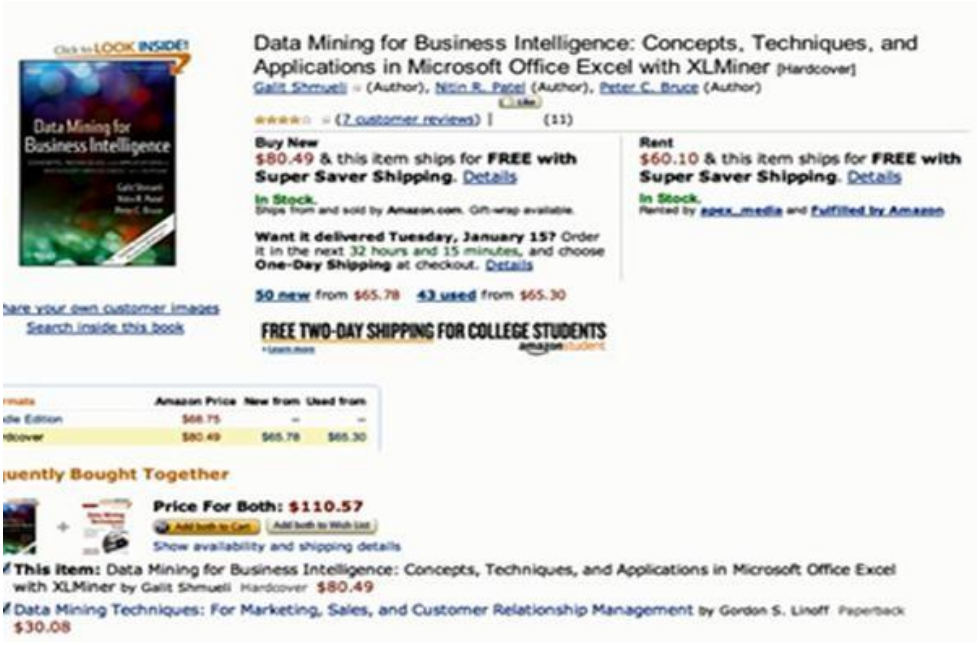
$$\frac{\text{Nombre de transactions contenant les produits X et Y}}{\text{Nombre total de transactions}}$$

  - **Confiance** : probabilité d'acheter le produit Y étant donné que le produit X a été acheté.

$$\frac{\text{Nombre de transactions contenant les produits X et Y}}{\text{Nombre de transactions contenant le produit X}}$$

## IV.1 - Règles d'association?

- Ce sont des règles de type:
  - **Si** le client achète le lait **alors** achète aussi le café
  - Notation: **Si** lait ➔ café
  - En général: Si antécédent ➔ conséquent
- Les règles d'association permet de:
  - trouver des combinaisons d'articles qui se produisent plus fréquemment dans une base de données transactionnelle
  - Mesurer la force et l'importance de ses combinaisons
  - Exemples?



## IV.2 - Représentation des transactions

- Nous pouvons représenter les transactions comme:
  - Liste
  - Représentation verticale
  - Représentation horizontale

Une liste

- Chaque ligne représente une transaction
- Chaque ligne liste les items achetés par le consommateur
- Les lignes peuvent avoir un numéro différent de colonnes

Liste de Items

| | A | B | C | D |
|---|---|---|---|---|
| 1 | | | | |
| 2 | tomates | lechuga | mostaza | jamon |
| 3 | tomates | pepinos | salad dressing | |
| 4 | agua | periodico | | |
| 5 | agua | coca-cola | | |
| 6 | | | | |
| 7 | | | | |

## IV.3 - Représentation verticale

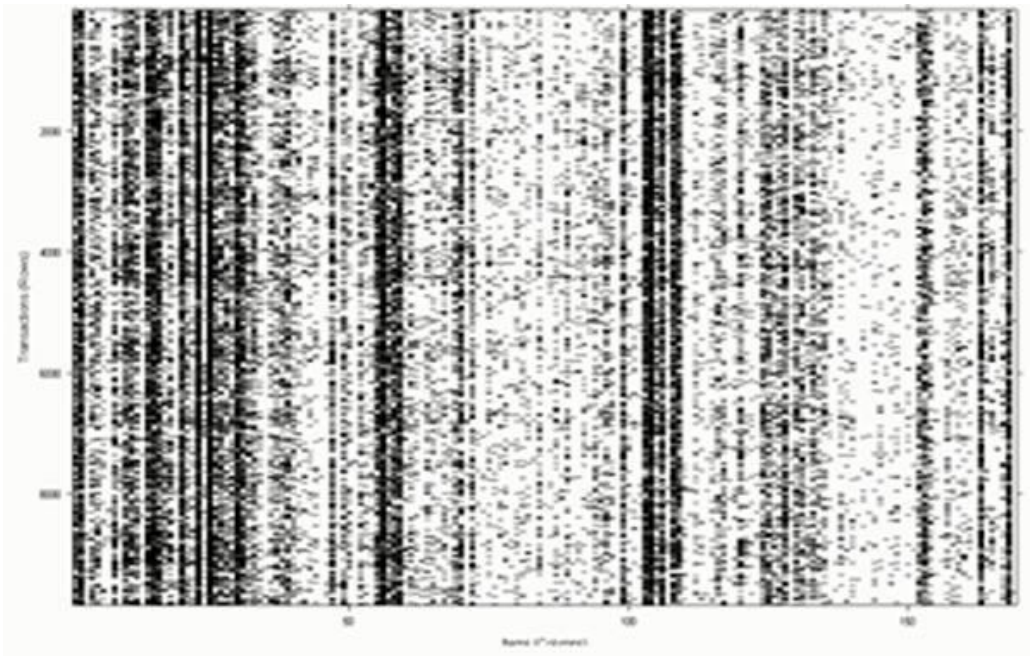
- Seulement deux colonnes
  - une colonne pour les numéros de la transaction (id)
  - Une colonne indiquant un item présent
- La forme mas efficace pour stocker les données

| A | B |
|---|---|
| TID | Item |
| 1 | tomates |
| 1 | lechuga |
| 1 | mostaza |
| 1 | jamon |
| 2 | tomates |
| 2 | pepinos |
| 2 | salad-dressing |
| 3 | agua |
| 3 | periodico |
| 4 | agua |
| 4 | coca-cola |

## IV.4 - Représentation horizontale

- Les transactions se représentent avec une matrice binaire :
  - Chaque ligne de la matrice représente une transaction
  - Chaque colonne représente un article ou item
  - Si un item est présent dans une transaction sera représenté avec un 1
  - Si un item est absent sera représenté avec un 0

| A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|
| TID | tomates | lechuga | mostaza | jamon | pepinos | salad-dressing | agua | periodico | coca-cola |
| 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |



## IV.5 - Critères d'évaluation des règles d'association

- Problème :
  - Agrawal (1994) découvre une méthode efficace pour trouver les règles
  - l'un des problèmes majeurs lorsque nous voulons traiter les règles d'association, c'est que nous pouvons trouver nombreuses (souvent trop) règles
  - Comment limiter le nombre des règles ? Comment rendre manipulable le processus de traitement postérieur ?
  - La réponse est dans les métriques que nous utilisons pour mesurer l'importance ou l'intérêt d'une règle.

## IV.6 - Métriques : Critères d'évaluation des règles d'association

- **SUPPORT** : un indicateur de « fiabilité » de la règle
- **CONFIANCE** : un indicateur de « précision » de la règle
- **LIFT** : Un indicateur de pertinence des règles

Dépasser le support et la confiance avec le

LIFT

Support

- Une règle donnée : « Si A➔B », le support de cette règle se définit comme le numéro de fois ou fréquence (relative) avec laquelle A et B figurent ensemble dans une base de données transactionnelle.
- Support peut être défini individuellement pour les items, mais aussi peut être défini pour la règle
- La première condition nous pouvons imposer pour limiter le nombre de règles est d'avoir un support minimum
- L'univers 1000 transaction
- Support (ordinateur)=400
- Support(ordinateur)= 400/1000

  = 0.4

Le support d'un ordinateur est la probabilité d'apparition d'un ordinateur dans une transaction


- Support(imprimante)=50
- Support(imprimante)=50/1000

  =0.05

P(imprimante)=0.05


- Support(Ordinateur et Imprimante)=40
- Support(Ordinateur et Imprimante)=40/1000= 0.04
- C'est la probabilité conjointe,

P(Ordinateur et imprimante)=0.04

Confiance

- Une règle données « Si A➔B », la confiance de cette règle est le quotient du support de la règle et le support de l'antécédent seulement.
- Confiance (A➔B)=support(A➔B)/support(A)
- Si le support mesure la fréquence, la confiance mesure la précision de la règle
- En langage de probabilité, la confiance est la probabilité conditionnelle:

Confiance (A➔B)=P(B/A)

**« Bonne » règle = règle avec un support et une confiance élevée**

Règles d'association

- **Support minimum σ :**
  - **Elevé** ⇒ **peu** d'itemsets fréquents
    ⇒ **peu** de règles valides qui ont été **souvent** vérifiées
  - **Réduit** ⇒ **plusieurs** règles valides qui ont été **rarement** vérifiées
- **Confiance minimum γ :**
  - **Elevée** ⇒ **peu** de règles, mais toutes "**pratiquement**" **correctes**
  - **Réduite** ⇒ **plusieurs** règles, plusieurs d'entre elles sont "**incertaines**"
- **Valeurs utilisées** : σ = 2 - 10 %, γ = 70 - 90 %

- La confiance(Imprimante ➔Ordinateur)=?
- La confiance(Ordinateur ➔ Imprimante)=?

La confiance(Imprimante ➔Ordinateur)=Support(Imprimante ➔Ordinateur) / support (Imprimante)

=40/50=0.8

La confiance(Ordinateur ➔ Imprimante)=Support(Ordinateur ➔ Imprimante) /support (Ordinateur)

=40/400=0.1

LIFT

| | Compro Pan | No Compro Pan | |
|---|---|---|---|
| Compro JN | 280 | 120 | 400 |
| No Compro JN | 420 | 180 | 600 |
| | 700 | 300 | 1000 |

Exercice

- Calculer:
  - Support(pain)
  - Support(Jus d'orange)
  - Support(pain➔jus d'orange)
  - Support(jus d'orange➔pain)
  - Confiance(pain➔ jus d'orange)
  - Confiance(jus d'orange➔pain)

Solution

  - Support(pain)=0.7
  - Support(Jus d'orange)=0.4
  - Support(pain➔jus d'orange)=0.28
  - Support(jus d'orange➔pain)=0.28
  - Confiance(pain➔ jus d'orange)=0.28/0.7=0.4
  - Confiance(jus d'orange➔pain)=0.28/0.4=0.7

Lift

- Est défini de la manière suivante:

Lift(A➔B)=support(A➔B)/(support(A)*support(B) )

- Lift=1 ou très proche de 1 indique que la relation est produite au hasard
- Lift supérieur à 1 traduit une corrélation positive de X et Y, et donc le caractère significatif de l'association
- Lift<1 indique une relation réellement faible

- Malheureusement n'existe pas de valeurs critiques pour déterminer c'est quoi « loin de 1 » ou au dessous de 1

- Pour un commerce au détail, le nombre de règles d'association possibles est souvent énorme. Vouloir étudier toutes les associations entre des produits à un niveau très fin de granularité amènerait à des résultats non interprétables. Pour obtenir des résultats cohérents et utiles, il faut tout d'abord faire une liste pertinente des règles d'association d'intérêt.

- Si le support est petit, il faut se questionner sur l'intérêt de la règle d'association. En pratique, on peut fixer un support minimum requis et exclure les règles d'association n'ayant pas le support requis.

- Un niveau de confiance très élevé ou très faible peut aussi gonfler (ou réduire) artificiellement le lift.

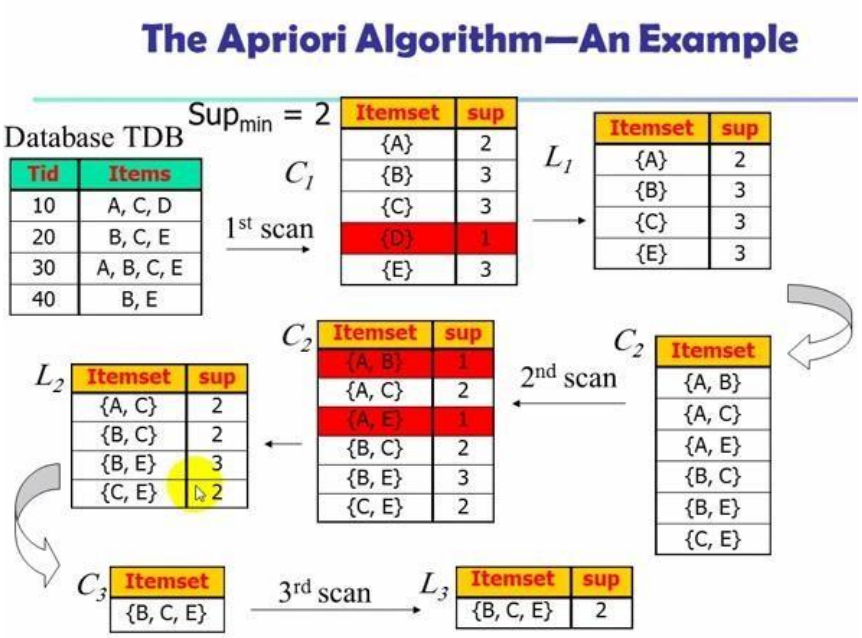
- L'objectif d'étudier les produits concomitants est de mieux comprendre une dynamique du comportement du client. En d'autres mots, on veut découvrir des associations non connues et prendre des décisions d'affaires basées sur ces nouvelles connaissances.

- Les règles qui obtiennent un bon support, une bonne confiance et un bon lift sont potentiellement utiles. Toutefois, ces règles peuvent être triviales, inexplicables ou difficiles à traduire en actions concrètes. Il faut au départ bien choisir les règles à étudier.

<u>Lift</u>

- Lift(pain➔Jus d'orange)=Lift(jus d'orange➔pain)=1
- Lift(Imprimate➔ordinateur)=lift(ordinateur➔imprimante)=0.004/(0.4*0.05)=2.00

## IV.7 - L'algorithme Apriori[Agrawal93]




### Extraction des règles d'association (I)

Démarche

Paramètres : Fixer un degré d'exigence sur les règles à extraire
>> Support min. (ex. 2 transactions)
>> Confiance min. (ex. 75%)

→ L'idée est surtout de contrôler (limiter) le nombre de règles produites

Démarche : Construction en deux temps
>> recherche des itemsets fréquents (support >= support min.)
>> à partir des itemsets fréquents, produire les règles (conf. >= conf. min.)

Quelques définitions :
>> item = produit
>> itemset = ensemble de produits (ex. {p1,p3})
>> sup(itemset) = nombre de transactions d'apparition simultanée des produits (ex. sup{p1,p3} = 4)
>> card(itemset) = nombre de produits dans l'ensemble (ex. card{p1,p3} = 2)

Exemple

## Extraction des Règles d'Association (II)

Recherche des Itemsets Fréquents

Cas général : $2^J$-1

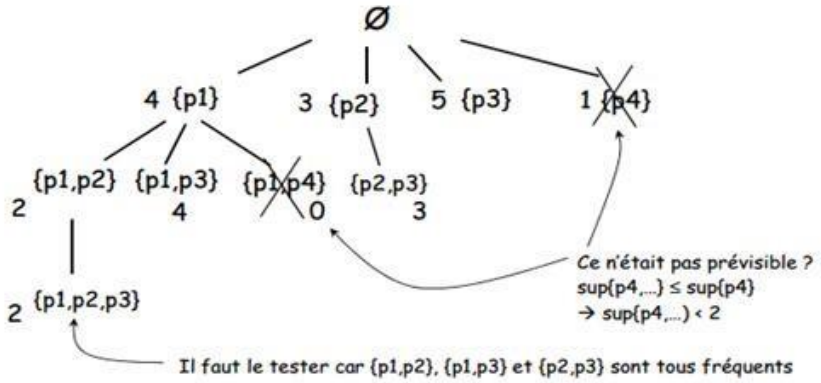
>> Nombre de calculs énormes
>> Chaque calcul impose de revenir scanner la base

$C_4^1 = 4$ ← Itemsets de card = 1
$C_4^2 = 6$ ← Itemsets de card = 2
$C_4^3 = 4$ ← Itemsets de card = 3
$C_4^4 = 1$
$\Sigma = 15 = 2^4 - 1$ ...

Réduire l'exploration en éliminant d'emblée certaines pistes

Données

| Caddie | p1 | p2 | p3 | p4 |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 |
| 2 | 1 | 0 | 1 | 0 |
| 3 | 1 | 1 | 1 | 0 |
| 4 | 1 | 0 | 1 | 0 |
| 5 | 0 | 1 | 1 | 0 |
| 6 | 0 | 0 | 0 | 1 |




Que se passerait-il si nous avions sup. min. = 3 ?

## Extraction des Règles d'Association (III)

Recherche des règles pour les itemsets de card = 2

Il faut tester toutes les combinaisons : 2 tests par itemset

Données

| Caddie | p1 | p2 | p3 | p4 |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 |
| 2 | 1 | 0 | 1 | 0 |
| 3 | 1 | 1 | 1 | 0 |
| 4 | 1 | 0 | 1 | 0 |
| 5 | 0 | 1 | 1 | 0 |
| 6 | 0 | 0 | 0 | 1 |

{p1,p2} { p1→p2 : conf. = 2/4 = 50% (refusé)
p2→p1 : conf. = 2/3 = 67% (refusé)

{p1,p3} { p1→p3 : conf. = 4/4 = 100% (accepté)
p3→p1 : conf. = 4/5 = 80% (accepté)

{p2,p3} { p2→p3 : conf. = 3/3 = 100% (accepté)
p3→p2 : conf. = 3/5 = 60% (refusé)

Que se passerait-il si nous avions conf. min. = 55 %

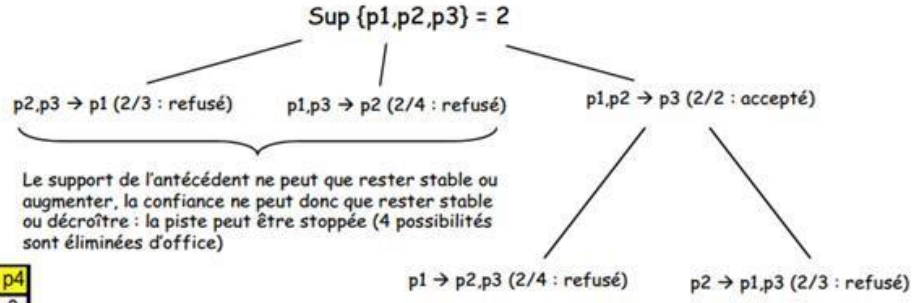
## Extraction des Règles d'Association (IV)

Recherche des règles pour les itemsets de card = 3 et plus...

$C_3^1 = 3$ ← Règles avec conséquent de card = 1

$C_3^2 = 3$ ← Règles avec conséquent de card = 2

Réduire l'exploration en éliminant d'emblée certaines pistes




Données

| Caddie | p1 | p2 | p3 | p4 |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 |
| 2 | 1 | 0 | 1 | 0 |
| 3 | 1 | 1 | 1 | 0 |
| 4 | 1 | 0 | 1 | 0 |
| 5 | 0 | 1 | 1 | 0 |
| 6 | 0 | 0 | 0 | 1 |

Que se passerait-il si nous avions conf. min. = 55 %

# Algorithme Apriori (Agrawal 93)

- Première passe :
  - recherche des 1-itemsets fréquents
  - un compteur par produits
- L'algorithme génère un candidat de taille k à partir de deux candidats de taille k-1 différents par le dernier élément
  - Apriori-Gen
- Passe k :
  - comptage des k-itemsets fréquents candidats
  - sélection des bons candidats

$C_k$: Itemset candidat de taille k
$L_k$ : itemset fréquent de taille k

$L_1$ = {fréquent items};
for ($k$ = 1; $L_k$ !=∅; $k$++) do begin
  $C_{k+1}$ = Apriori-Gen (candidats générés à partir de $L_k$)
  for each transaction $t$ dans la base do
    ***incrémenter le nombre de candidats dans $C_{k+1}$ qui sont dans t***
  $L_{k+1}$ = candidats dans $C_{k+1}$ avec un support_min
end
return $\cup_k L_k$;

There are lots of measures proposed in the literature

Some measures are good for certain applications, but not for others

What criteria should we use to determine whether a measure is good or bad?

What about Apriori-style support based pruning? How does it affect these measures?

| # | Measure | Formula |
|---|---|---|
| 1 | $\phi$-coefficient | $\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$ |
| 2 | Goodman-Kruskal's ($\lambda$) | $\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$ |
| 3 | Odds ratio ($\alpha$) | $\frac{P(A,B)P(\overline{A},\overline{B})}{P(A,\overline{B})P(\overline{A},B)}$ |
| 4 | Yule's $Q$ | $\frac{P(A,B)P(\overline{AB})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{AB})+P(A,\overline{B})P(\overline{A},B)} = \frac{\alpha-1}{\alpha+1}$ |
| 5 | Yule's $Y$ | $\frac{\sqrt{P(A,B)P(\overline{AB})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{AB})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$ |
| 6 | Kappa ($\kappa$) | $\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$ |
| 7 | Mutual Information ($M$) | $\frac{\sum_i \sum_j P(A_i,B_j)\log \frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i)\log P(A_i), -\sum_j P(B_j)\log P(B_j))}$ |
| 8 | J-Measure ($J$) | $\max\left(P(A,B)\log(\frac{P(B\|A)}{P(B)}) + P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),\right.$ $\left. P(A,B)\log(\frac{P(A\|B)}{P(A)}) + P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\right)$ |
| 9 | Gini index ($G$) | $\max\Big(P(A)[P(B\|A)^2 + P(\overline{B}\|A)^2] + P(\overline{A})[P(B\|\overline{A})^2 + P(\overline{B}\|\overline{A})^2]$ $-P(B)^2 - P(\overline{B})^2,$ $P(B)[P(A\|B)^2 + P(\overline{A}\|B)^2] + P(\overline{B})[P(A\|\overline{B})^2 + P(\overline{A}\|\overline{B})^2]$ $-P(A)^2 - P(\overline{A})^2\Big)$ |
| 10 | Support ($s$) | $P(A,B)$ |
| 11 | Confidence ($c$) | $\max(P(B\|A), P(A\|B))$ |
| 12 | Laplace ($L$) | $\max\left(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2}\right)$ |
| 13 | Conviction ($V$) | $\max\left(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})}\right)$ |
| 14 | Interest ($I$) | $\frac{P(A,B)}{P(A)P(B)}$ |
| 15 | cosine ($IS$) | $\frac{P(A,B)}{\sqrt{P(A)P(B)}}$ |
| 16 | Piatetsky-Shapiro's ($PS$) | $P(A,B) - P(A)P(B)$ |
| 17 | Certainty factor ($F$) | $\max\left(\frac{P(B\|A)-P(B)}{1-P(B)}, \frac{P(A\|B)-P(A)}{1-P(A)}\right)$ |
| 18 | Added Value ($AV$) | $\max(P(B\|A) - P(B), P(A\|B) - P(A))$ |
| 19 | Collective strength ($S$) | $\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})} \times \frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$ |
| 20 | Jaccard ($\zeta$) | $\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$ |
| 21 | Klosgen ($K$) | $\sqrt{P(A,B)}\max(P(B\|A) - P(B), P(A\|B) - P(A))$ |

# Chapitre 4
# Clustering

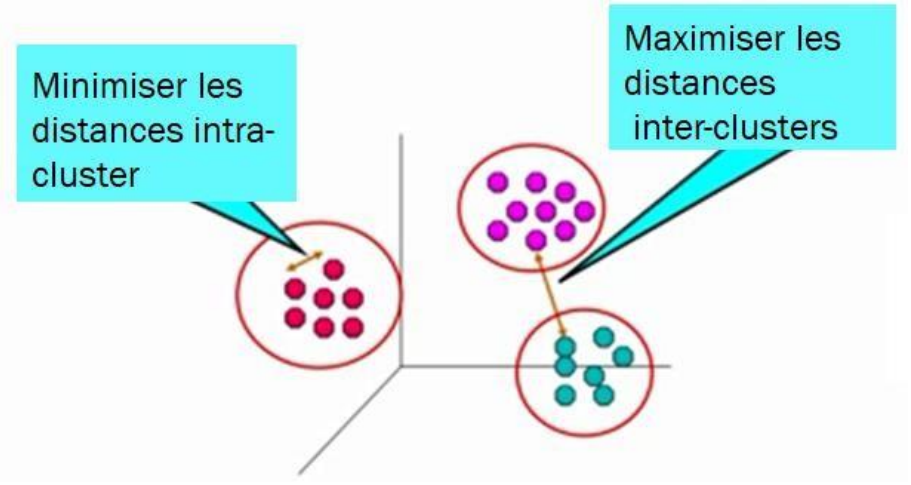
# V) Clustering (Segmentation)

- Soient N instances de données à n attributs,
- Trouver un partitionnement en k clusters (groupes) ayant un sens (Similitude)
- Affectation automatique de “labels” aux clusters
- k peut être donné, ou “découvert”
- Plus difficile que la classification car les classes ne sont pas connues à l’avance (non supervisé)
- Attributs
  - Numériques (distance bien définie)
  - Enumératifs ou mixtes (distance difficile à définir)

## V.1 - Qualité d’un clustering

- Une bonne méthode de clustering produira des clusters d’excellente qualité avec :
  - Similarité importante **intra-classe**
  - Similarité faible **inter-classe**
- La **qualité** d’un clustering dépend de :
  - La mesure de similarité utilisée
  - L’implémentation de la mesure de similarité
- La **qualité d’une méthode** de clustering est évaluée par son habilité à découvrir certains ou tous les “patterns” cachés

## V.2 - Objectifs du clustering



## V.3 - Exemples d'applications

- **Marketing :** segmentation du marché en découvrant des groupes de clients distincts à partir de bases de données d'achats.
- **Environnement :** identification des zones terrestres similaires (en termes d'utilisation) dans une base de données d'observation de la terre.
- **Assurance:** identification de groupes d'assurés distincts associés à un nombre important de déclarations.
- **Planification de villes:** identification de groupes d'habitations suivant le type d'habitation, valeur, localisation géographique, …
- **Médecine**: Localisation de tumeurs dans le cerveau
    - Nuage de points du cerveau fournis par le neurologue
    - Identification des points définissant une tumeur

## V.4 - Méthodes de Clustering

- Méthode de partitionnement(K-Means)
- Méthodes hiérarchiques (par agglomération)
- Méthode des voisinages denses
- Caractéristique:
    - Apprentissage non supervisé (classes inconnues)
    - Toutes les variables ont le même statut
        - Pas de variable dépendante (prédictive)
    - Pb: interprétation des clusters identifiés

Méthode K-means

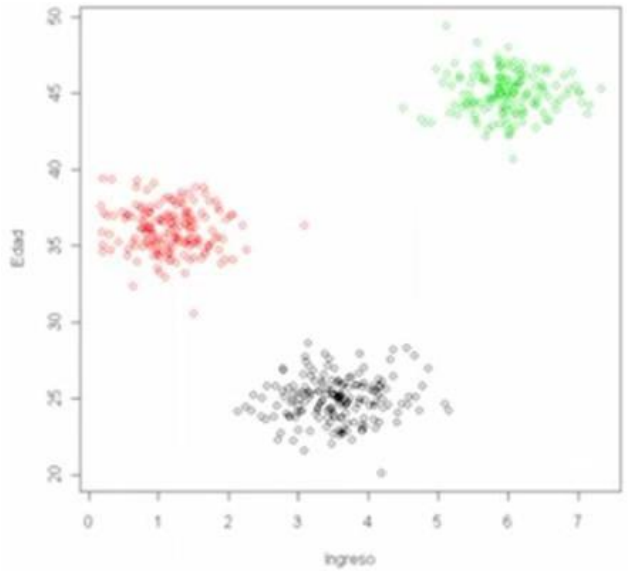
- K-Means est une méthode clustering
- Les observations d'un groupe doivent être similaires aux autres observations du groupe, mais …
    - Doivent être différentes des observations de autres groupes.

Concepts basiques de K-Means

- techniquement, nous voulons maximiser la variation inter-cluster et minimiser la variation intra-cluster
- Exemple:
    - Nous avons des données des âges et revenus pour un groupe de consommateurs.

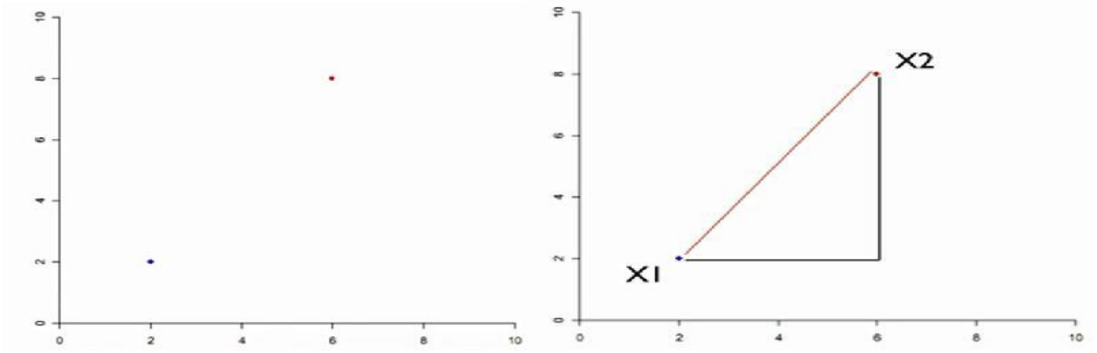
- La question? Existe-ils des groupes de consommateurs avec des caractéristiques similaires dans cette base de données?
- Dans cet exemple simple avec deux variables nous pouvons représenter graphiquement les données

Combien de groupes?




Distance

- «Observations similaires »➔
- « des observations que son proches»
- Ça veut dire quoi proche?
  - Nous avons besoin du concept de la distance pour pouvoir parler de proche et loin




Distance

- Nous pouvons utiliser la mesure que nous avons apprise en primaire (théorème de Pythagore)

- Techniquement connue comme ***distance euclidienne.***
- La distance entre le point X1(2,2) et le point x2(6,8) égale à

$$D(x_1, x_2) = [\ (6-2)^2 + (8-2)^2\ ]^{0.5} = 7.21$$

- En général, la distance euclidienne est définie pour deux vecteurs de p dimensions (variables)

$$d(\mathbf{X}_i, \mathbf{X}_j) = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2}$$

- Pour les variables continues on va utiliser cette distance (existe d'autres distances)

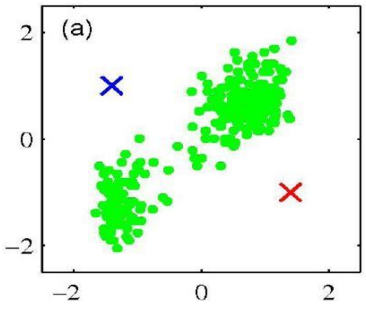
<u>L'algorithme K-Means</u>

- K-Means
  - Cette méthode suppose que nous connaissons le numéro de groupes(clusters)
    - Donc la méthode trouve la « **meilleure** » affectation de points aux différents groupes (clusters)
    - « la **meilleure** » dans le sens de maximiser la distance inter-clusters et minimiser la distance intra-cluster

<u>Algorithme</u>

- Décider le numéro (#) de clusters. Notons k à ce numéro
- Une méthode possible d'initialisation: prendre K observations aléatoirement des données. Ces observations se deviennent les K centres initiaux c1,c2,…,ck.
- Pour chaque N-K observations restantes, calculons les distance entre l'observation correspondante et chacun des centres
- Chaque observation est alors affectée au centre le plus proche
- À la fin de l'affectation des observations nous aurons K groupes d'observations.
- Pour chacun de ces groupes, Nous calculons les nouveaux centres. Le centre est un vecteur des moyennes pour toutes les variables utilisées par les observations au sein de chaque groupe.
- Répéter le processus ….
- jusqu'à ce qu'il n'y ai plus de réaffectation

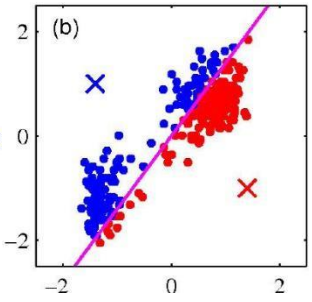
## K-means clustering: Example




- Pick *K* random points as cluster centers (means)

Shown here for *K*=2

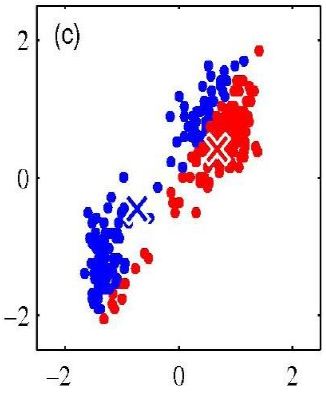
## K-means clustering: Example




Iterative Step 1

- Assign data points to closest cluster center

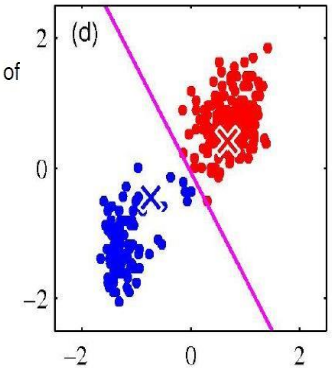
## K-means clustering: Example




Iterative Step 2

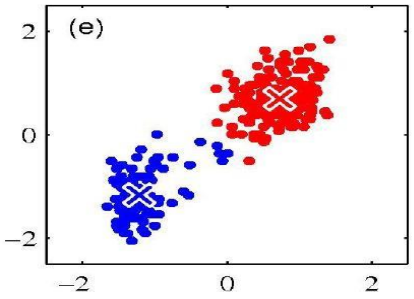
- Change the cluster center to the average of the assigned points

## K-means clustering: Example

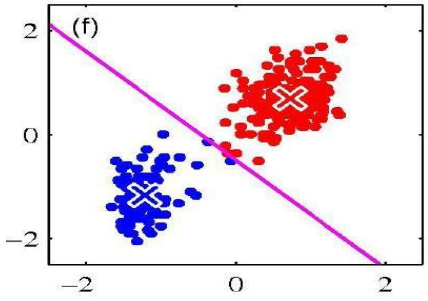


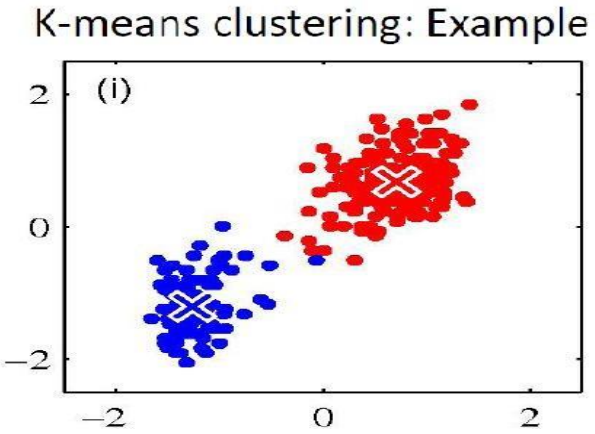

- Repeat until convergence

## K-means clustering: Example




## K-means clustering: Example






Exemple

| | | Var1 | Var2 | Var3 | Var4 | Var5 | Distances Cluster1 | Cluster2 | Cluster3 | Minimo | ClusterAsignado |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | 7 | 8 | 4 | 5 | 2 | 7.14 | 9.00 | 5.20 | 5.20 | 3 |
| | B | 6 | 8 | 5 | 4 | 2 | 7.07 | 8.49 | 4.90 | 4.90 | 3 |
| | C | 8 | 9 | 7 | 8 | 9 | 3.16 | 9.38 | 9.49 | 3.16 | 1 |
| | D | 6 | 7 | 7 | 7 | 8 | 0.00 | 6.32 | 6.93 | 0.00 | 1 |
| | E | 1 | 2 | 5 | 3 | 4 | 9.27 | 4.24 | 4.90 | 4.24 | 2 |
| | F | 3 | 4 | 5 | 3 | 5 | 6.06 | 3.61 | 3.87 | 3.61 | 2 |
| | G | 7 | 8 | 8 | 6 | 6 | 2.83 | 7.62 | 7.07 | 2.83 | 1 |
| | H | 8 | 9 | 6 | 5 | 5 | 4.69 | 8.83 | 7.07 | 4.69 | 1 |
| | I | 2 | 3 | 5 | 6 | 5 | 6.78 | 2.83 | 3.16 | 2.83 | 2 |
| | J | 1 | 2 | 4 | 4 | 2 | 10.20 | 5.83 | 4.24 | 4.24 | 3 |
| | K | 3 | 2 | 6 | 5 | 7 | 6.32 | 0.00 | 5.48 | 0.00 | 2 |
| | L | 2 | 5 | 6 | 8 | 9 | 4.80 | 4.80 | 6.71 | 4.80 | 3 |
| | M | 3 | 5 | 4 | 6 | 3 | 6.93 | 5.48 | 0.00 | 0.00 | 3 |
| | N | 3 | 5 | 5 | 6 | 3 | 6.56 | 5.20 | 1.00 | 1.00 | 3 |
| Centroids | | | | | | | | | | | |
| Cluster1 | D | 6 | 7 | 7 | 7 | 8 | | | | | |
| Cluster2 | K | 3 | 2 | 6 | 5 | 7 | | | | | |
| Cluster3 | M | 3 | 5 | 4 | 6 | 3 | | | | | |

| | | Var1 | Var2 | Var3 | Var4 | Var5 | Distances Cluster1 | Cluster2 | Cluster3 | Minimo | ClusterAsignado |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | 7 | 8 | 4 | 5 | 2 | 6.41 | 7.93 | 3.88 | 3.88 | 3 |
| | B | 6 | 8 | 5 | 4 | 2 | 6.36 | 7.23 | 3.36 | 3.36 | 3 |
| | C | 8 | 9 | 7 | 8 | 9 | 3.04 | 10.16 | 9.32 | 3.04 | 1 |
| | D | 6 | 7 | 7 | 7 | 8 | 0.92 | 7.09 | 6.93 | 0.92 | 1 |
| | E | 1 | 2 | 5 | 3 | 4 | 9.36 | 2.30 | 5.37 | 2.30 | 2 |
| | F | 3 | 4 | 5 | 3 | 5 | 6.83 | 1.95 | 3.83 | 1.95 | 2 |
| | G | 7 | 8 | 8 | 6 | 6 | 2.20 | 7.83 | 6.46 | 2.20 | 1 |
| | H | 8 | 9 | 6 | 5 | 5 | 3.85 | 8.56 | 6.07 | 3.85 | 1 |
| | I | 2 | 3 | 5 | 6 | 5 | 6.96 | 1.82 | 4.95 | 1.82 | 2 |
| | J | 1 | 2 | 4 | 4 | 2 | 10.16 | 3.78 | 4.82 | 3.78 | 3 |
| | K | 3 | 2 | 6 | 5 | 7 | 6.76 | 2.30 | 6.14 | 2.30 | 2 |
| | L | 2 | 5 | 6 | 8 | 9 | 5.39 | 5.81 | 7.71 | 5.39 | 1 |
| | M | 3 | 5 | 4 | 6 | 3 | 6.70 | 3.91 | 1.70 | 1.70 | 3 |
| | N | 3 | 5 | 5 | 6 | 3 | 6.34 | 3.72 | 1.75 | 1.75 | 3 |
| Centroids | | | | | | | | | | | |
| Cluster1 | | 6.2 | 7.6 | 6.8 | 6.8 | 7.4 | | | | | |
| Cluster2 | | 2.25 | 2.75 | 5.25 | 4.25 | 5.25 | | | | | |
| Cluster3 | | 4 | 5.6 | 4.4 | 5 | 2.4 | | | | | |

Attention !

- Pas de garantie que l'algorithme trouve la solution optimale
- Une mauvaise sélection initiale des centres peut conduire à un groupement pauvre
- Recommandation: Exécuter l'algorithme plusieurs fois avec des points différents.
- K-means, comme n'importe quel algorithme qui se calcule à base des distances, peut être affecté par les unités de mesure des variables
    - Les variables mesurées en grandes unités dominent la construction des clusters

- Recommandation: Standardiser les variables avant de commencer la recherche des clusters.

Avantages de K-Means

- Rapidité, peut être appliqué à des bases données relativement grandes
- Economique de point de vue stockage de données (stoker les K centres)

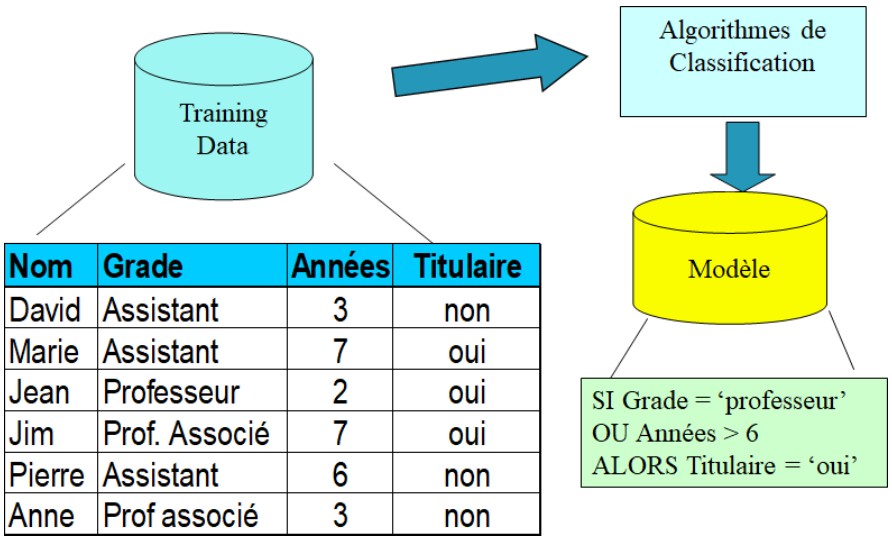
Inconvénients K-means

- Suppose la connaissance de K (en réalité jamais connu)
- Sensible à la présence des observations extrêmes

# Chapitre 5
# Classification

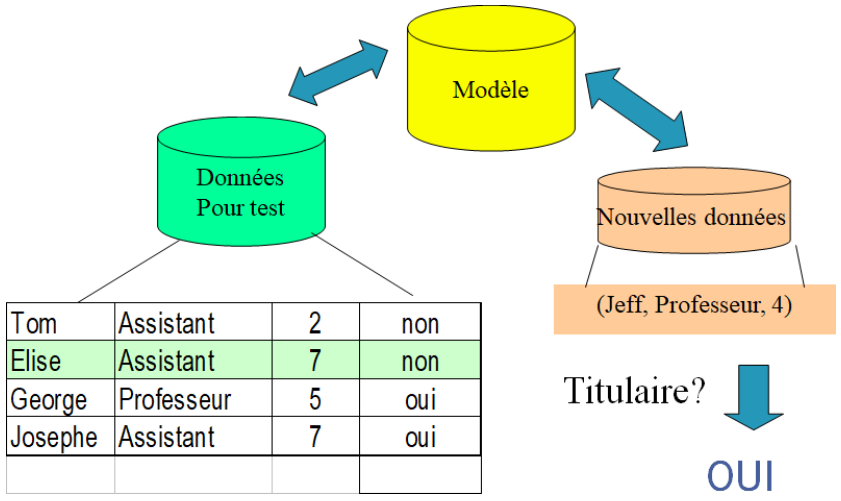
# VI) Classification Datamining: Méthodes prédictives

## VI.1 - Arbre de décision méthode de classification

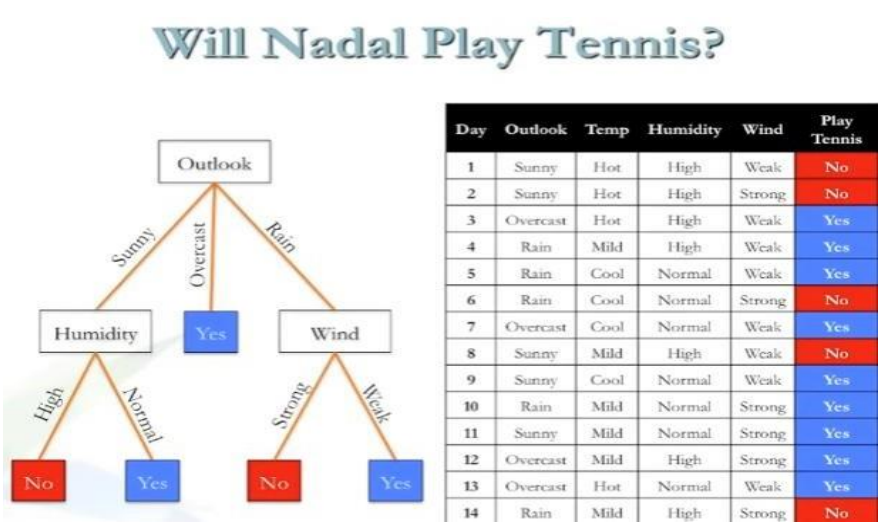



| Nom | Grade | Années | Titulaire |
|---|---|---|---|
| David | Assistant | 3 | non |
| Marie | Assistant | 7 | oui |
| Jean | Professeur | 2 | oui |
| Jim | Prof. Associé | 7 | oui |
| Pierre | Assistant | 6 | non |
| Anne | Prof associé | 3 | non |

### VI.1.1) Processus de Classification (2): Prédiction




| | | | |
|---|---|---|---|
| Tom | Assistant | 2 | non |
| Elise | Assistant | 7 | non |
| George | Professeur | 5 | oui |
| Josephe | Assistant | 7 | oui |
| | | | |


quelles sont les variables à utiliser et dans quel ordre? Quel critère utilisé pour sélectionner la «meilleure» division?

Les mesures **d'impureté** suivantes seront utilisées: l'erreur de classification, l'indice de GINI et l'entropie, pour cela se définie la probabilité:

***p(j/t)*** =la probabilité d'appartenance à la classe « j » étant dans le Nœud t.

Souvent notée par $p_j$

**Question #2: l'erreur de classification, l'indice de Gini et l'entropie seront utilisés**

- Erreur de classification:

$$Error(t) = 1 - \max_j [\, p(j \mid t)]$$

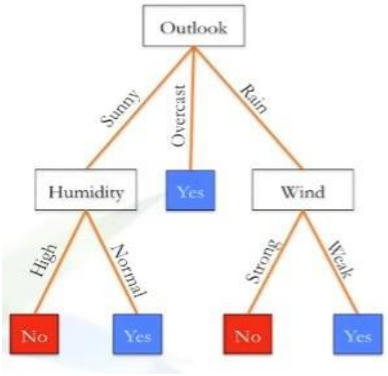
- Indice de Gini

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

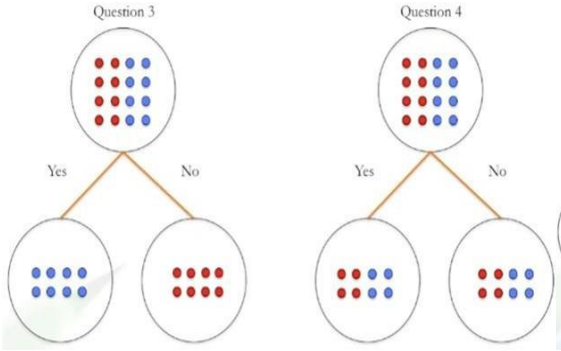
- Entropie:

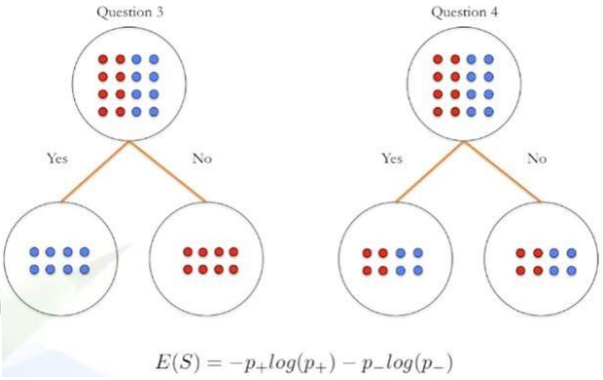
$$Entropia(t) = -\sum_j p(j \mid t) \log_2 p(j \mid t)$$

## Will Nadal Play Tennis?

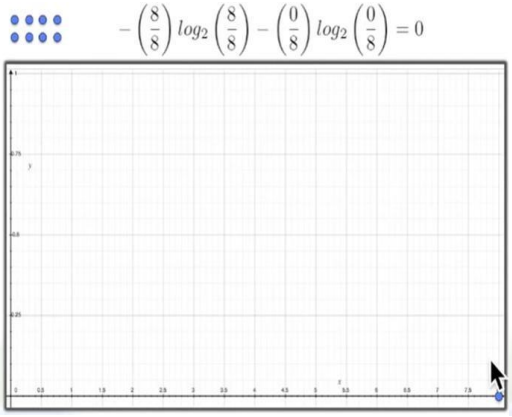



| Day | Outlook | Temp | Humidity | Wind | Play Tennis |
|---|---|---|---|---|---|
| 1 | Sunny | Hot | High | Weak | No |
| 2 | Sunny | Hot | High | Strong | No |
| 3 | Overcast | Hot | High | Weak | Yes |
| 4 | Rain | Mild | High | Weak | Yes |
| 5 | Rain | Cool | Normal | Weak | Yes |
| 6 | Rain | Cool | Normal | Strong | No |
| 7 | Overcast | Cool | Normal | Weak | Yes |
| 8 | Sunny | Mild | High | Weak | No |
| 9 | Sunny | Cool | Normal | Weak | Yes |
| 10 | Rain | Mild | Normal | Strong | Yes |
| 11 | Sunny | Mild | Normal | Strong | Yes |
| 12 | Overcast | Mild | High | Strong | Yes |
| 13 | Overcast | Hot | Normal | Weak | Yes |
| 14 | Rain | Mild | High | Strong | No |

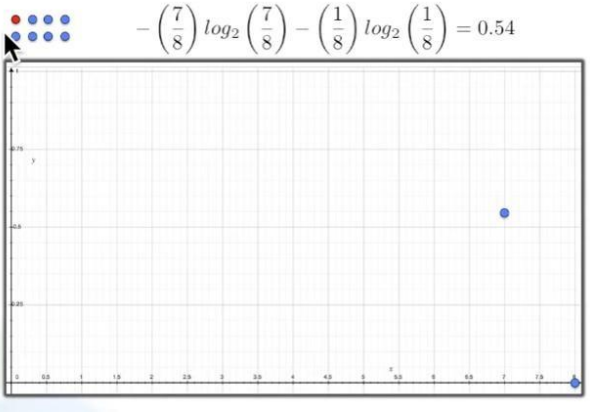


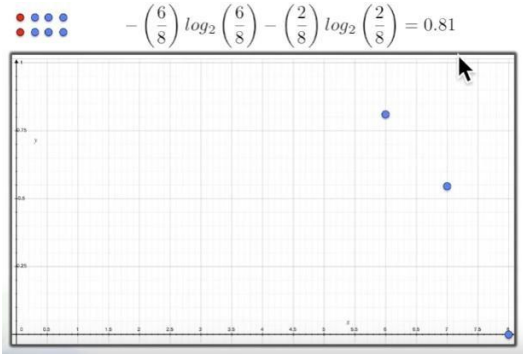

$$E(S) = -p_+ log(p_+) - p_- log(p_-)$$

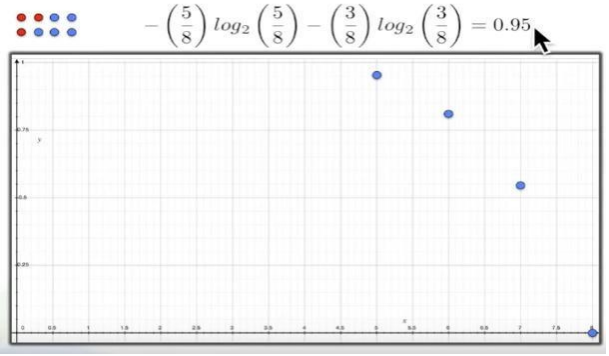


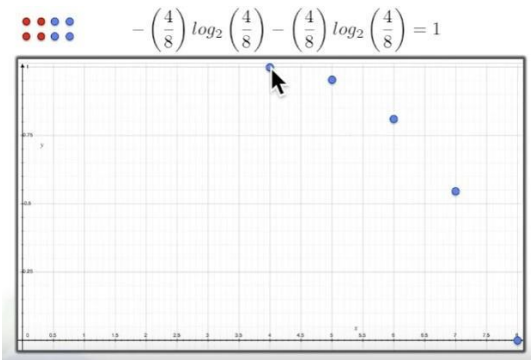
$$-\left(\frac{8}{8}\right) log_2 \left(\frac{8}{8}\right) - \left(\frac{0}{8}\right) log_2 \left(\frac{0}{8}\right) = 0$$

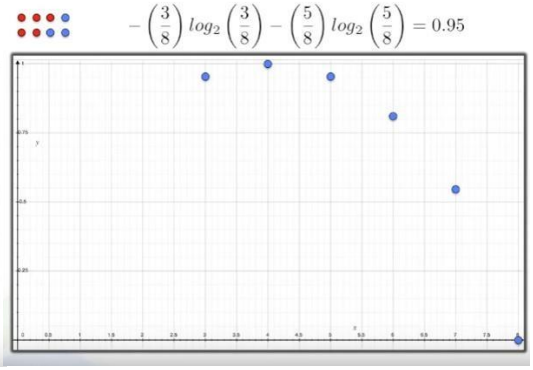


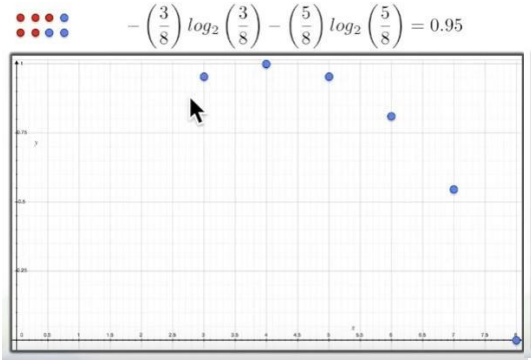
$$-\left(\frac{7}{8}\right) log_2 \left(\frac{7}{8}\right) - \left(\frac{1}{8}\right) log_2 \left(\frac{1}{8}\right) = 0.54$$

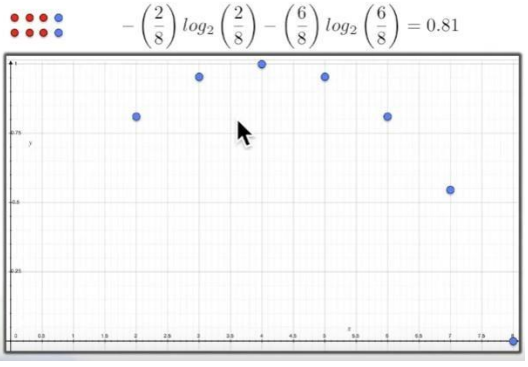

$-\left(\frac{6}{8}\right) log_2\left(\frac{6}{8}\right) - \left(\frac{2}{8}\right) log_2\left(\frac{2}{8}\right) = 0.81$

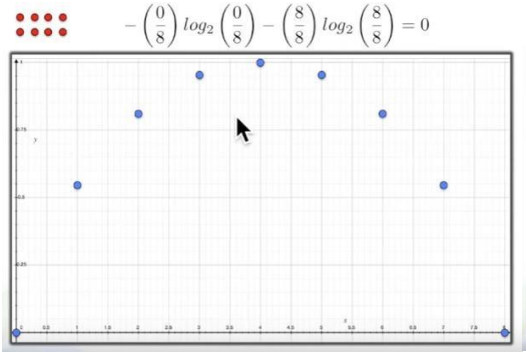

$-\left(\frac{5}{8}\right) log_2\left(\frac{5}{8}\right) - \left(\frac{3}{8}\right) log_2\left(\frac{3}{8}\right) = 0.95$

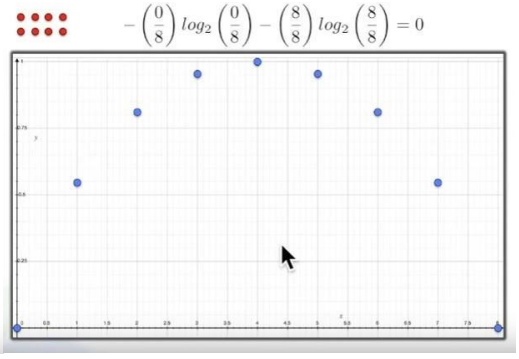

$-\left(\frac{4}{8}\right) log_2\left(\frac{4}{8}\right) - \left(\frac{4}{8}\right) log_2\left(\frac{4}{8}\right) = 1$

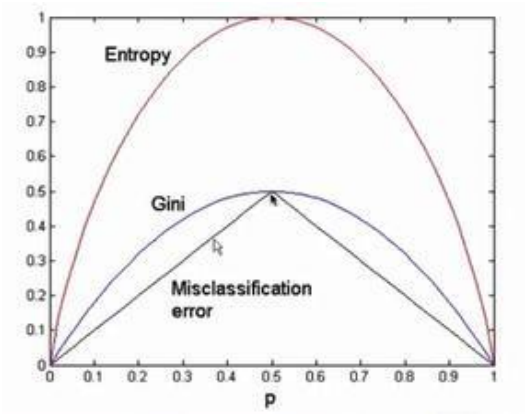

$-\left(\frac{3}{8}\right) log_2\left(\frac{3}{8}\right) - \left(\frac{5}{8}\right) log_2\left(\frac{5}{8}\right) = 0.95$


$-\left(\frac{3}{8}\right) log_2\left(\frac{3}{8}\right) - \left(\frac{5}{8}\right) log_2\left(\frac{5}{8}\right) = 0.95$


$-\left(\frac{2}{8}\right) log_2\left(\frac{2}{8}\right) - \left(\frac{6}{8}\right) log_2\left(\frac{6}{8}\right) = 0.81$

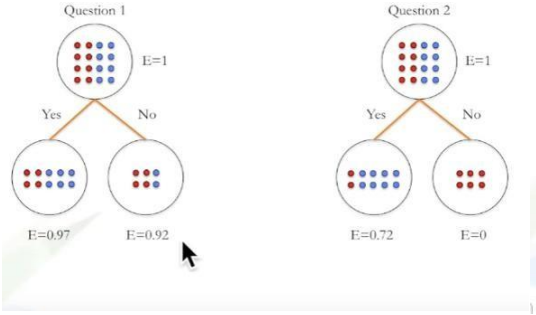

$-\left(\frac{0}{8}\right) log_2\left(\frac{0}{8}\right) - \left(\frac{8}{8}\right) log_2\left(\frac{8}{8}\right) = 0$

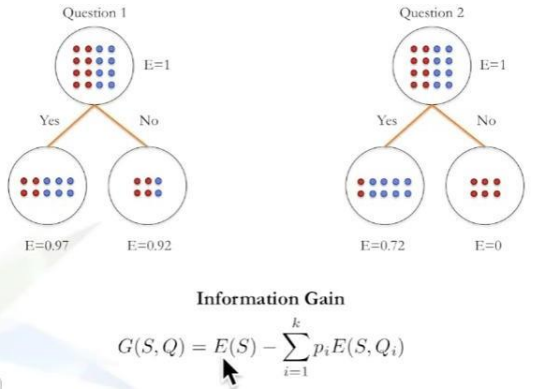

$-\left(\frac{0}{8}\right) log_2\left(\frac{0}{8}\right) - \left(\frac{8}{8}\right) log_2\left(\frac{8}{8}\right) = 0$




$$y = -\sum_{i=1}^{k} p_i log_k(p_i)$$

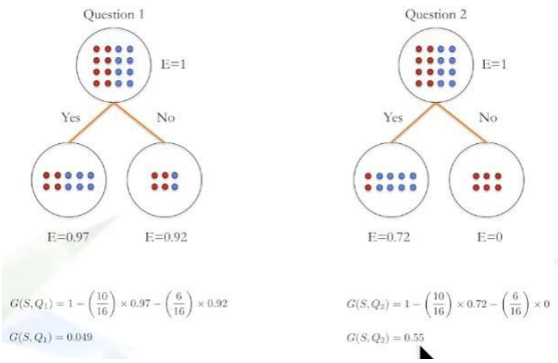
$$y = -\left[\left(\frac{1}{10}\right)log_4\left(\frac{1}{10}\right)\right] - \left[\left(\frac{3}{10}\right)log_4\left(\frac{3}{10}\right)\right] - \left[\left(\frac{2}{10}\right)log_4\left(\frac{2}{10}\right)\right] - \left[\left(\frac{4}{10}\right)log_4\left(\frac{4}{10}\right)\right]$$




**Information Gain**

$$G(S,Q) = E(S) - \sum_{i=1}^{k} p_i E(S,Q_i)$$

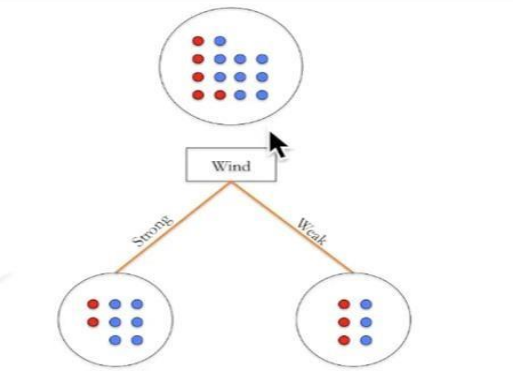


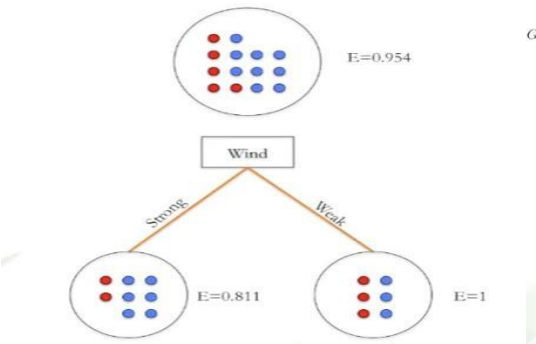

$G(S,Q_1) = 1 - \left(\frac{10}{16}\right) \times 0.97 - \left(\frac{6}{16}\right) \times 0.92$

$G(S,Q_1) = 0.049$

$G(S,Q_2) = 1 - \left(\frac{10}{16}\right) \times 0.72 - \left(\frac{6}{16}\right) \times 0$

$G(S,Q_2) = 0.55$

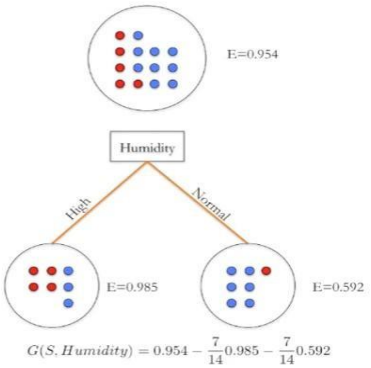



$G(S,Wind) = 0.954 - \frac{8}{14}0.811 - \frac{6}{14}1$

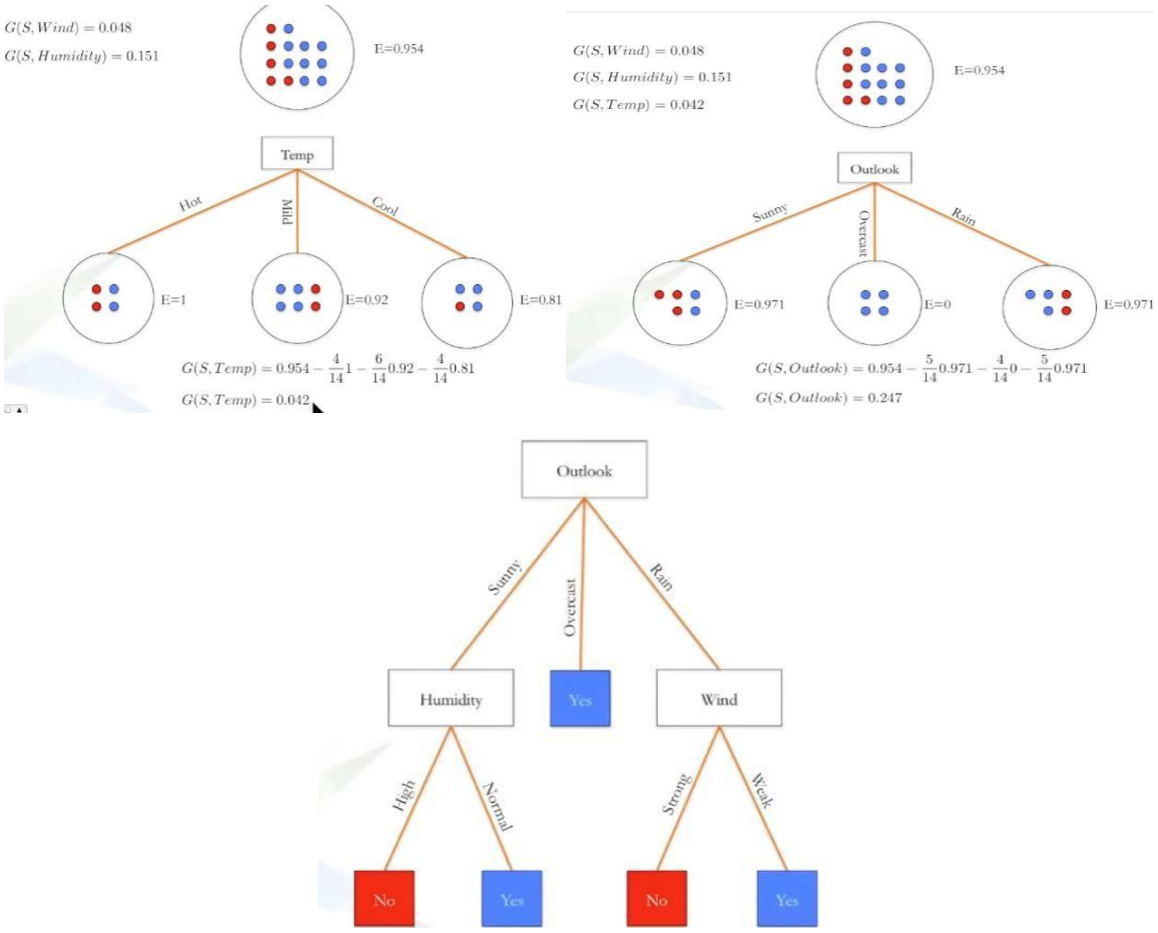
$G(S,Wind) = 0.048$

$G(S,Wind) = 0.048$



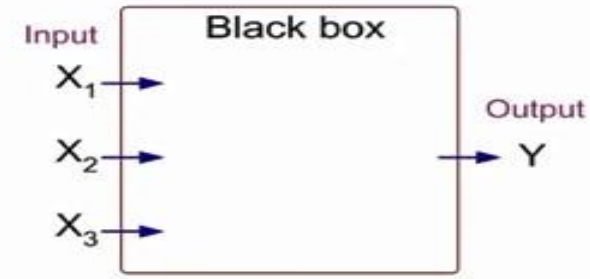

$G(S,Humidity) = 0.954 - \frac{7}{14}0.985 - \frac{7}{14}0.592$

$G(S,Humidity) = 0.151$



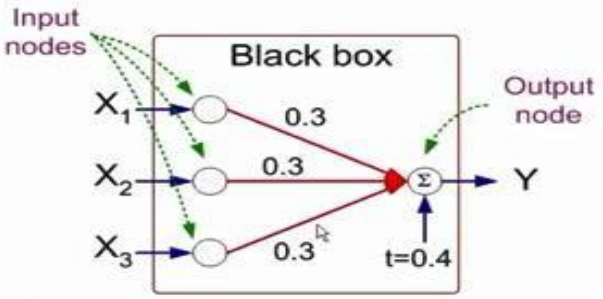

## VI.2 - Classification Réseaux de Neurones

| $X_1$ | $X_2$ | $X_3$ | Y |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 1 | 0 | 1 | 1 |
| 1 | 1 | 0 | 1 |
| 1 | 1 | 1 | 1 |
| 0 | 0 | 1 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 1 | 1 | 1 |
| 0 | 0 | 0 | 0 |




La sortie Y est 1 si au moins 2 de 3 entrés sont égales à 1

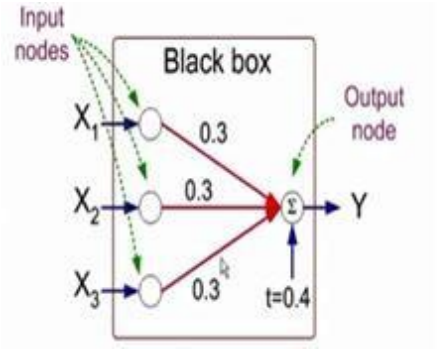
| $X_1$ | $X_2$ | $X_3$ | Y |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 1 | 0 | 1 | 1 |
| 1 | 1 | 0 | 1 |
| 1 | 1 | 1 | 1 |
| 0 | 0 | 1 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 1 | 1 | 1 |
| 0 | 0 | 0 | 0 |




$$Y = I(0.3X_1 + 0.3X_2 + 0.3X_3 > 0.4)$$

$$\text{donde } I(z) = \begin{cases} 1 & \text{Si } z \text{ es verdadero} \\ 0 & \text{En otro caso} \end{cases}$$

- Un réseau de neurones est composé de plusieurs neurones interconnectés. Un poids est associé à chaque arc. A chaque neurone on associe une valeur
- Le Nœud de sortie est la somme pondérée des valeurs de sorties des neurones
- Comparer le nœud de sortie avec un **seuil t**

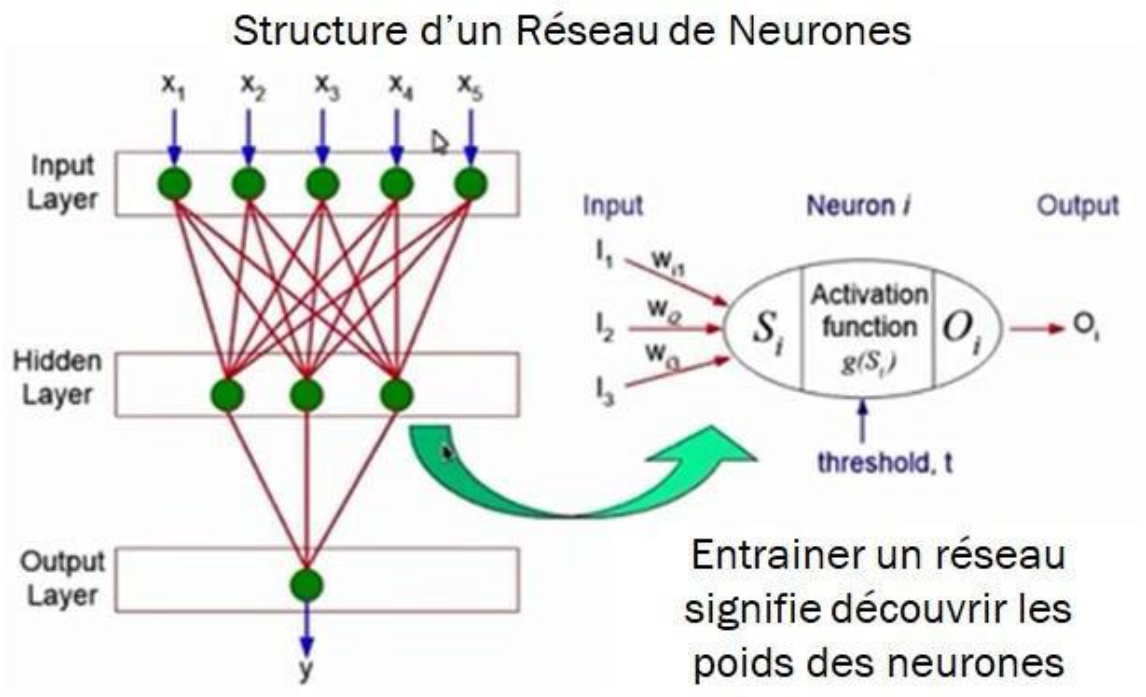



Modèle perceptron

$$Y = I(\sum_i w_i X_i - t)$$

$$Y = sign(\sum_i w_i X_i - t)$$

## VI.3 - Structure d'un Réseau de Neurones




### VI.3.1) Algorithme d'apprentissage

- Initialiser les poids ($w_1, w_2, \ldots, w_k$)
- Ajuster les poids de sorte que la sortie du réseau de neurones soit en accord avec les étiquettes des classes d'entrainement.
- Fonction objective:

$$E = \sum_i [Y_i - f(w_i, X_i)]^2$$

  - Trouver les poids $w'_i$ qui minimise la fonction objective antérieure (erreur quadratique)
  - Un critère d'arrêt doit être défini
  - Exemple: **Backpropagation**

## Apprentissage : « Back propagation »

1er étape : Initialiser les poids des liens entre les neurones. Souvent *une valeur entre 0 et 1*, déterminée aléatoirement, est assignée à chacun des poids.

2e étape : Application d'un vecteur entrées-sorties à apprendre.

3e étape : Calcul des sorties du RNA à partir des entrées qui lui sont appliquées et calcul de l'erreur entre ces sorties et les sorties idéales à apprendre.

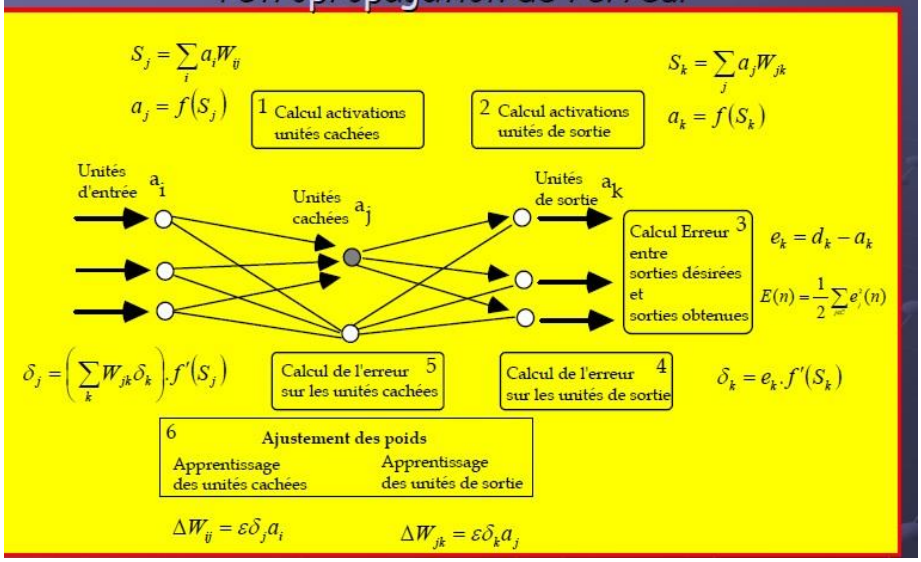
4e étape : Correction des poids des liens entre les neurones de la couche de sortie et de la première couche cachée selon l'erreur présente en sortie.

5e étape : Propagation de l'erreur sur la couche précédente et correction des poids des liens entre les neurones de la couche cachée et ceux en entrées.

6e étape : Boucler à la 2e étape avec un nouveau vecteur d'entrées-sorties tant que les performances du RNA (erreur sur les sorties) ne sont pas satisfaisantes.

59

## Le perceptron multicouche apprentissage : retropropagation de l'erreur



# Chapitre 6
# Régression

# VII) Régression

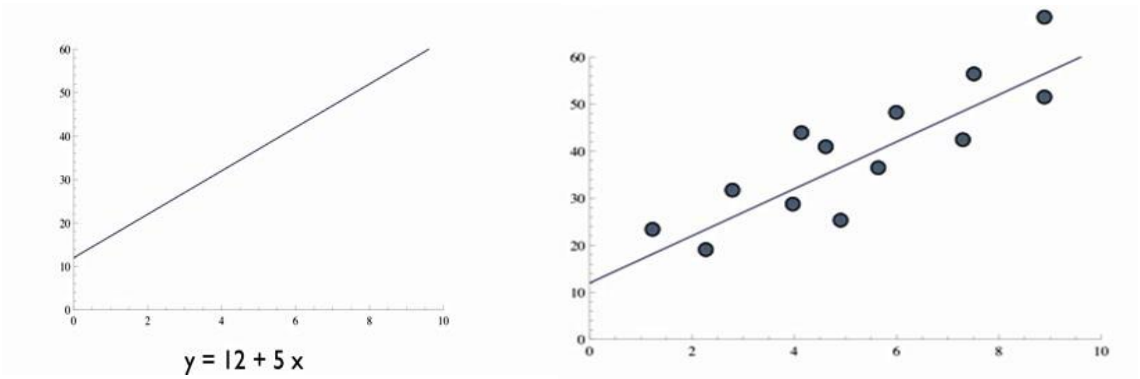
## VII.1 - Régression Linéaire simple

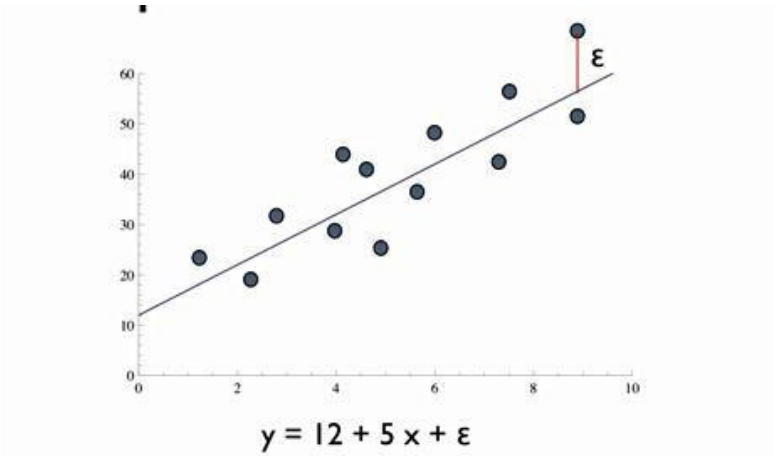
- Nous poursuivons deux objectifs:
  1. Etablir s'il y a une relation/corrélation entre deux variable
     - Existe –elle une relation statistique significative entre la consommation et le revenu?
  2. Prédire des nouvelles observations
     - Combien seront les ventes d'un produit dans les prochaines 4 mois ?

Régression simple

- Supposons que nous avons deux types de variables
  - Une variable dépendante Y que nous voulons expliquer ou prédire.
  - Une variable indépendante X que explique la variable Y.
- On va supposer que les deux variable sont connectées à travers une équation linéaire.

Une équation linéaire Y=b+aX




L'équation de régression complète

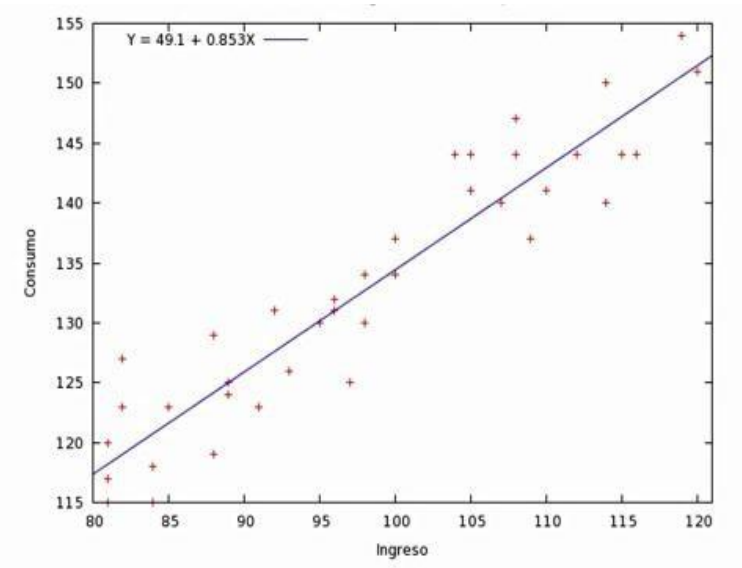
- Y=b+a*X+ε
- Exemple:
  - Nous avons les données de 40 familles sur la consommation d'un produit donnée (nourriture par exemple)
  - Nous avons aussi les données sur le revenu de ces familles

| | | | | | |
|---|---|---|---|---|---|
| 1 | 119 | 154 | 21 | 116 | 144 |
| 2 | 85 | 123 | 22 | 115 | 144 |
| 3 | 97 | 125 | 23 | 93 | 126 |
| 4 | 95 | 130 | 24 | 105 | 141 |
| 5 | 120 | 151 | 25 | 89 | 124 |
| 6 | 92 | 131 | 26 | 104 | 144 |
| 7 | 105 | 141 | 27 | 108 | 144 |
| 8 | 110 | 141 | 28 | 88 | 129 |
| 9 | 98 | 130 | 29 | 109 | 137 |
| 10 | 98 | 134 | 30 | 112 | 144 |
| 11 | 81 | 115 | 31 | 96 | 132 |
| 12 | 81 | 117 | 32 | 89 | 125 |
| 13 | 91 | 123 | 33 | 93 | 126 |
| 14 | 105 | 144 | 34 | 114 | 140 |
| 15 | 100 | 137 | 35 | 81 | 120 |
| 16 | 107 | 140 | 36 | 84 | 118 |
| 17 | 82 | 123 | 37 | 88 | 119 |
| 18 | 84 | 115 | 38 | 96 | 131 |
| 19 | 100 | 134 | 39 | 82 | 127 |
| 20 | 108 | 147 | 40 | 114 | 150 |

Consommation=49.1334+0.852736*revenu+ ε

Interprétation des Coefficients

- Coefficient de la constante:
  - Le 49.1334 signifie le niveau de consommation d'une famille avec un niveau de revenu=0.
  - constante n'a pas toujours une interprétation intuitive.
  - c'est clair parce que pas toujours un sens de parler d'une situation dans laquelle la variable indépendante est égale à zéro.
- Coefficient a (pente):
  - Le 0.852537 signifie que quand le revenu augmente d'une unité (un dollar par exemple) la consommation augmente 85 centimes de dollar.
  - Si je donne un dollar de plus à une famille sa consommation va augmenter 85 centimes de dollar le reste sera économiser ou achat d'autre produit.
  - Mesure la sensibilité de la variable dépendante Y à un changement un.




### VII.1.1) Prévision avec la régression linéaire simple

- maintenant une nouvelle famille apparaît (numéro 41)
- Nous avons l'information sur son revenu mais pas sur sa consommation.
- Son niveau de revenu est 100
- Pouvons prédire combien va consommer cette famille en utilisant la régression estimée?

Premier essai

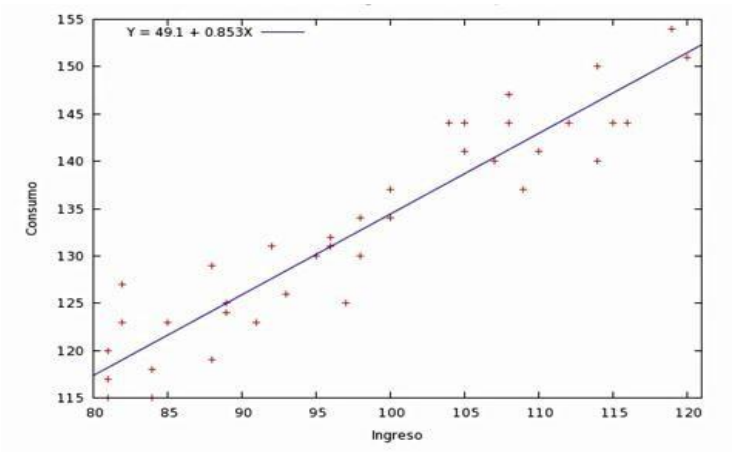
- On va remplacer la valeur du revenu=100 dans l'équation estimée:

**Consommation=49.1334+0.852736*revenu+ ε**

**Consommation=49.1334+0.852736*100+0**

**=134.407**

- Il s'agit d'une estimation ponctuelle (ponctuelle de point)
- Mais …

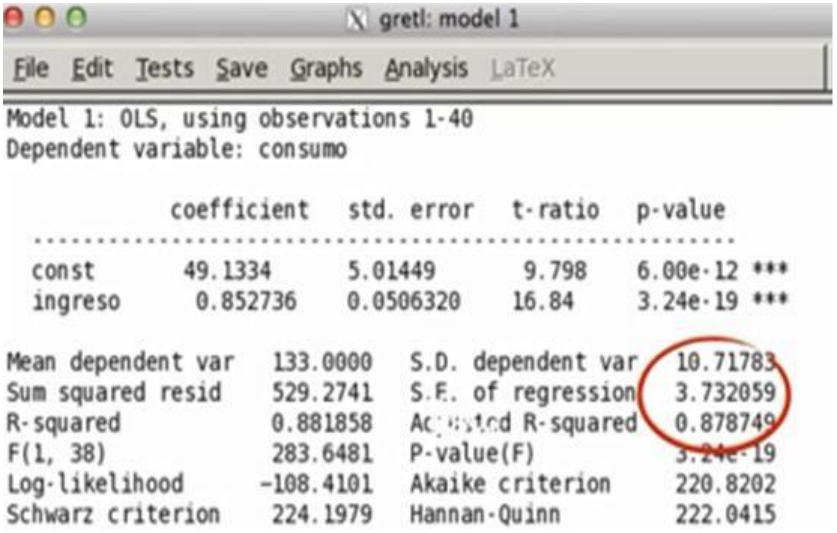


Deuxième Essai

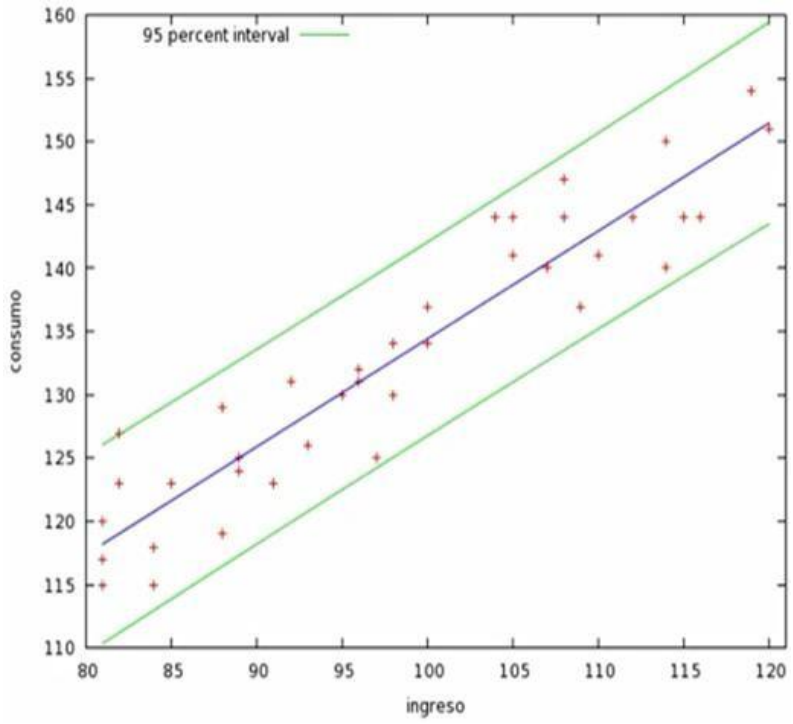
- Nous voulions construire un prévisionnel qui tient compte de la variabilité qui existe autour de la ligne de régression.
- Nous voulions au lieu d'un point un ensemble de valeurs possibles.
- Nous voulions pouvoir assigner un degré de confiance à de prévisionnel.
- Par exemple, nous voulions un prévisionnel (pronostique) avec un niveau de confiance de 95%
- Comment faire pour construire un intervalle de confiance pour ce pronostique (un de 95% de confiance)
- Facile:

**Le rang prévisionnel =pronostique ponctuel+/- 2*e.s.r**

**(e.s.r erreur type de la régresion)**




- Dans notre cas, notre intervalle pour le pronostique sera :
- 134. 4077+(-) 2*3.732059
- [126.94, 141.87]
- Ce pronostique est meilleur que le ponctuel de 134.4077
- Toujours quand je demande un pronostique je veux dire un pronostique pour rang.




Existe-elle une relation entre la consommation et le revenu?

- A partir de l'équation **Consommation=49.1334+0.852736*revenu**

Il semble que la réponse doit être si

- Cependant, nous nous précipitons un peu dans cette confirmation
- Nous devons prendre en considération que cette régression a été générée à partir d'un échantillon d'une population que nous intéresse.
- Où il y a des échantillons impliqués existe toujours une variabilité

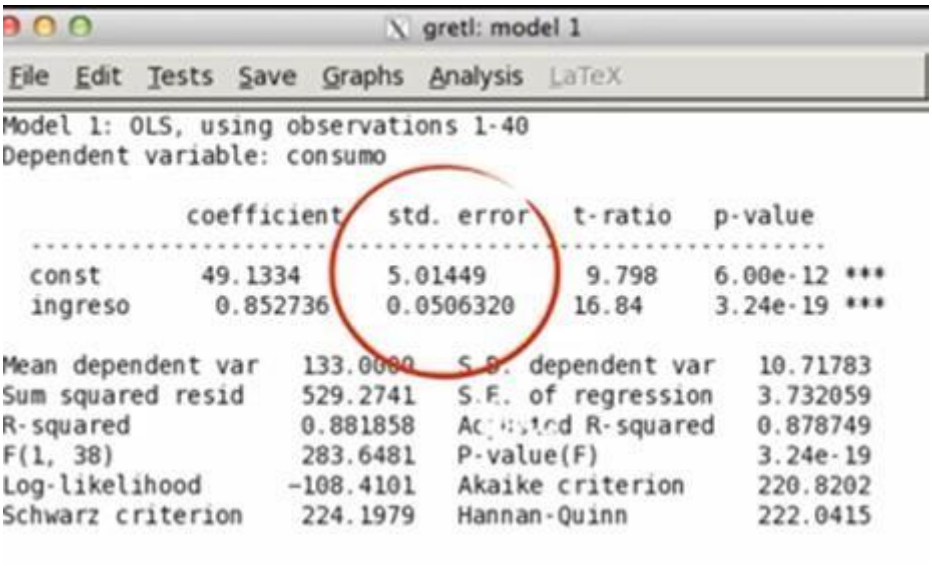
Intervalles de confiance

- Si nous travaillons avec des échantillons, toujours nous aurons besoins d'une bonde de confiance autour de nos valeurs estimées des coefficients d'une régression.
- Ces bondes sont les intervalles de confiance
- Les intervalles de confiance sont toujours associées a un niveau de confiance (typiquement 95%)

### VII.1.2) Comment construire un intervalle de confiance?

- Un intervalle de confiance de 95% pour un coefficient d'une régression (soit la constante ou la pente) se construit en sommant/restant à notre estimation de ce coefficient, 2 erreurs standards de ce coefficient.
- Ces erreurs standards on peut les retrouver dans le fichier de sortie des application statistiques à coté des coefficients respectivement.

Intervalles de confiance




- Intervalle de confiance pour la pente de la régression de notre exemple est donné par:

0.852736+/-2*0.0506320

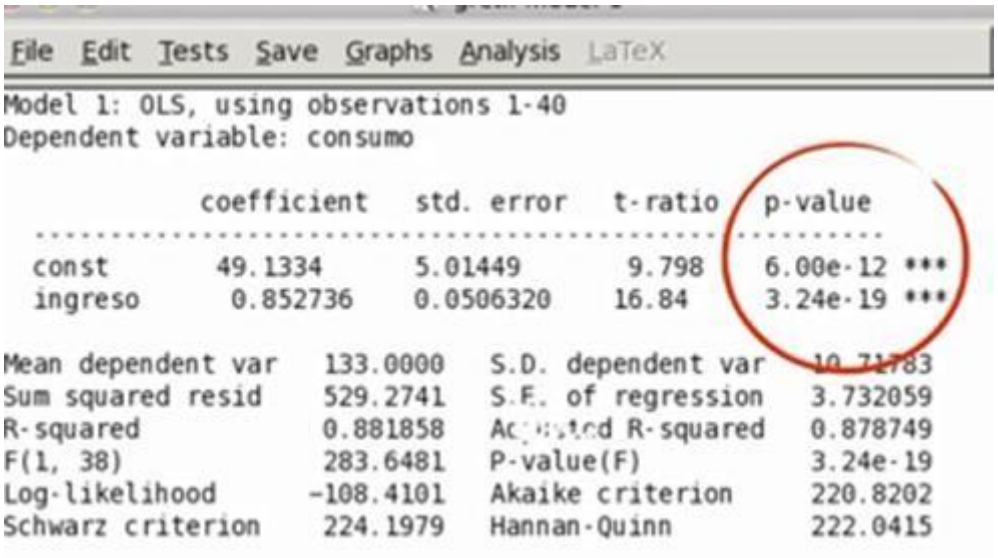
[0.7514,0.9540]

Toutes les valeurs de cet intervalle de pente sont cohérentes avec ce jeu de données.

- Comment utiliser un intervalle de confiance pour décider si la relation entre la variable (consommation et revenu) est statistiquement significative (à 95%) ?
- nous vérifions si la valeur de 0 est contenue dans cet intervalle
- Règle :
  - Si le 0 n'est pas contenu, la relation est statistiquement significative
  - Si le 0 est contenu, ne nous pouvons pas rejeter l'hypothèse que n'existe pas une la relation est statistiquement significative
  - 9a veut dire quoi une pente de 0 ?
  - Y=b+a*X➔ signifie que X pas d'impact sur Y (donc pas de relation entre X et Y)
- Si 0 est dans l'intervalle, les données sont cohérente avec l'hypothèse « pas de relation entre X et Y »
- Nous voulons rejeter cette hypothèse alternative (qui dit le contraire : existe une relation)
- Dans notre exemple l'intervalle [0.7514,0.9540] exclus le 0. Donc existe une relation.

<u>Autre Règle</u>

- Existe une autre règle (plus facile, mais moins intuitive) pour voir si il existe une relation statistiquement significative entre Y et X.
- Nous vérifions le p-value

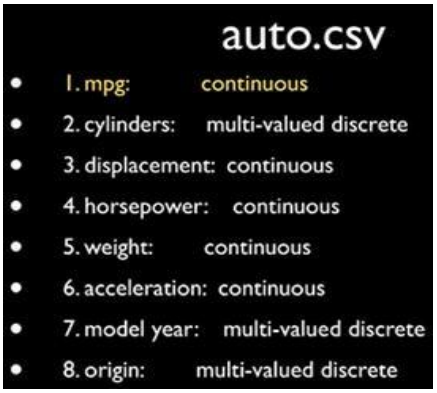



- Règle : Pour déterminer s'il existe une relation statistiquement significative entre la consommation et le revenue (à 95% de confiance) nous vérifions si la valeur-p de la pente est inférieure de 5% (ou 0.05).
- Si elle est supérieure à 0.05 nous ne pouvons pas rejeter l'hypothèse qu'il n'est pas de relation (95% de confiance)
- Notre Exemple :
- La p-value de la pente est 3.24e- 19 0.000000000000000000324<0.05 ➔ la relation est statistiquement significative

## VII.2 - Régression linéaire multiple

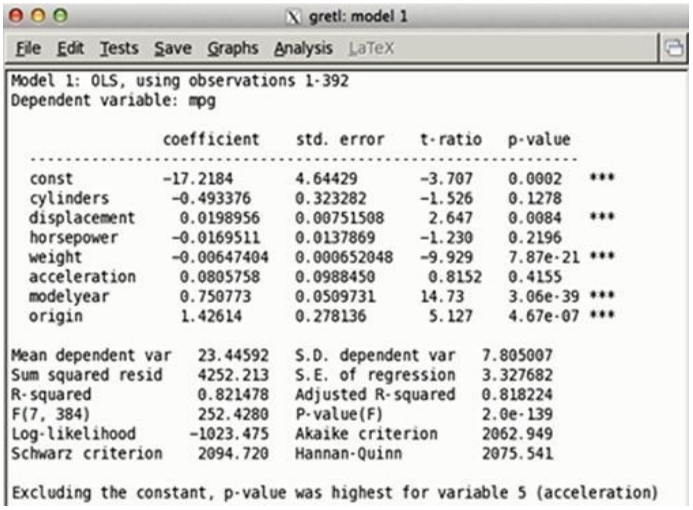
Sélection de variables

- Dans cette session on va discuter deux stratégies de sélection de modèle d'analyse de régression.
  - La première est basée sur la relation que existe entre la variable dépendante et chacune des variables indépendantes.
  - La deuxième est basée sur la contribution de chaque variable dans la prédiction de la variable dépendante.



Première stratégie

- Populaire
  - Consiste à éliminer les variables indépendantes **que ne sont pas statistiquement significative**
  - Si le niveau critique de signification est de 5%, cela équivalent à éliminer toutes les variables avec p-value supérieures à 5%
  - Le résultat une régression propre



gretl: model 1

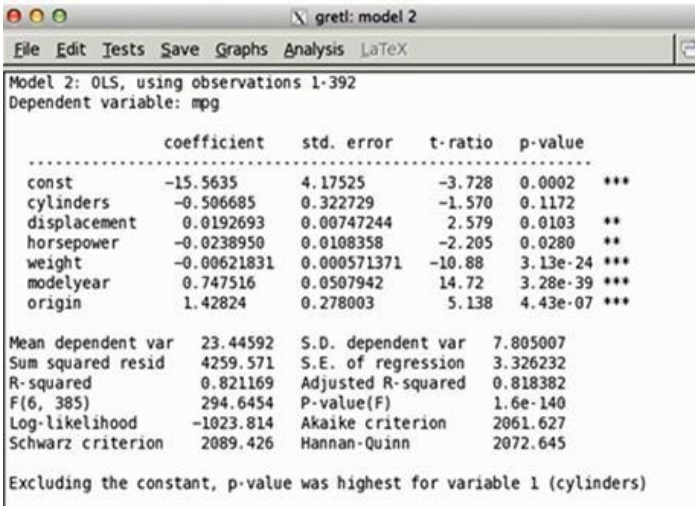
File Edit Tests Save Graphs Analysis LaTeX

```
Model 1: OLS, using observations 1-392
Dependent variable: mpg

                coefficient    std. error    t-ratio   p-value
  ------------------------------------------------------------
  const         -17.2184       4.64429      -3.707    0.0002   ***
  cylinders      -0.493376     0.323282     -1.526    0.1278
  displacement    0.0198956    0.00751508    2.647    0.0084   ***
  horsepower     -0.0169511    0.0137869    -1.230    0.2196
  weight         -0.00647404   0.000652048  -9.929    7.87e-21 ***
  acceleration    0.0805758    0.0988450     0.8152   0.4155
  modelyear       0.750773     0.0509731    14.73     3.06e-39 ***
  origin          1.42614      0.278136      5.127    4.67e-07 ***

Mean dependent var   23.44592   S.D. dependent var   7.805007
Sum squared resid    4252.213   S.E. of regression   3.327682
R-squared            0.821478   Adjusted R-squared   0.818224
F(7, 384)            252.4280   P-value(F)           2.0e-139
Log-likelihood      -1023.475   Akaike criterion     2062.949
Schwarz criterion    2094.720   Hannan-Quinn         2075.541

Excluding the constant, p-value was highest for variable 5 (acceleration)
```



gretl: model 2

File Edit Tests Save Graphs Analysis LaTeX

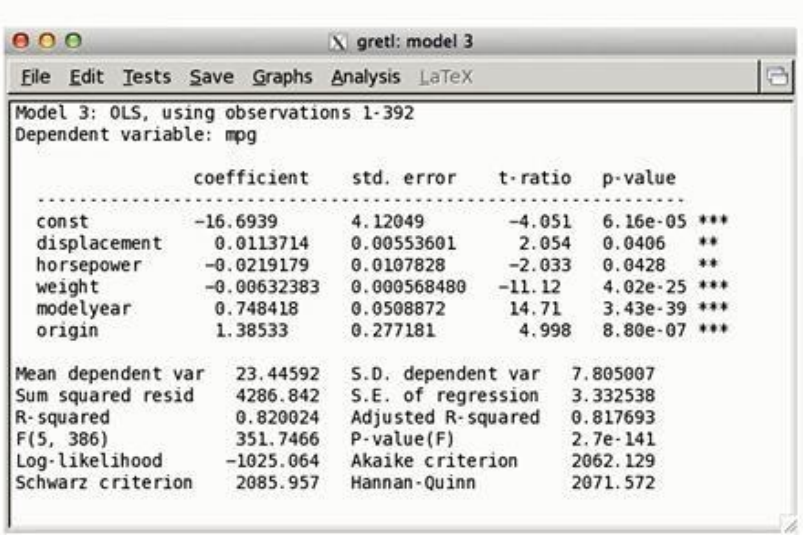
```
Model 2: OLS, using observations 1-392
Dependent variable: mpg

                coefficient    std. error    t-ratio   p-value
  ------------------------------------------------------------
  const         -15.5635       4.17525      -3.728    0.0002   ***
  cylinders      -0.506685     0.322729     -1.570    0.1172
  displacement    0.0192693    0.00747244    2.579    0.0103   **
  horsepower     -0.0238950    0.0108358    -2.205    0.0280   **
  weight         -0.00621831   0.000571371  -10.88    3.13e-24 ***
  modelyear       0.747516     0.0507942    14.72     3.28e-39 ***
  origin          1.42824      0.278003      5.138    4.43e-07 ***

Mean dependent var   23.44592   S.D. dependent var   7.805007
Sum squared resid    4259.571   S.E. of regression   3.326232
R-squared            0.821169   Adjusted R-squared   0.818382
F(6, 385)            294.6454   P-value(F)           1.6e-140
Log-likelihood      -1023.814   Akaike criterion     2061.627
Schwarz criterion    2089.426   Hannan-Quinn         2072.645

Excluding the constant, p-value was highest for variable 1 (cylinders)
```

- Eliminer un par un et recalculer p-value (dans cas acceleration)
- Horsepower mnt <0.05 par contre Cylinders encore >0.05 ➔ elimination de cylinders mais de Horsepower




- Toutes les variables avec p-value<0.05

Le problème de cette stratégie est que le critère de sélection est strictement strict.

Parfois en utilisant cette stratégie nous éliminons des informations pour pronostiquer la variable dépendante.

La définition de « statistiquement significative » est arbitraire. Personne ne peut confirmer qu'une p-value de 0.049 est beaucoup meilleure que 0.051.

Comment mesurer quand une variable contribue pour faire de bon pronostique ?

Nous pouvons utiliser le R-quadratique?

- Un premier candidat pour sélectionner des variables quand le pronostique nous intéresse est le R-quadratique
- Après tout nous disons que le R-quadratique mesure la qualité d'ajustement.
- Plus de qualité d'ajustement la prévision est meilleur?

Problèmes avec R-quadratique

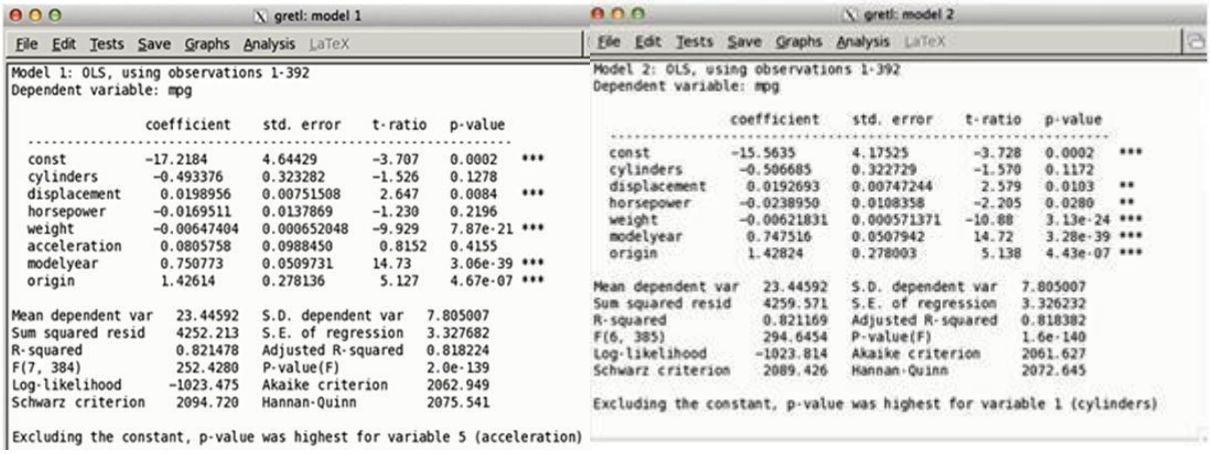
- Le R-quadratique toujours augmente en ajoutant de nouvelle variable dans la régression
- Augmente aussi quand la variable ajouter est absurde
- Si nous utilisons ce critère toujours on va sélectionner le modèle avec plus de variable.

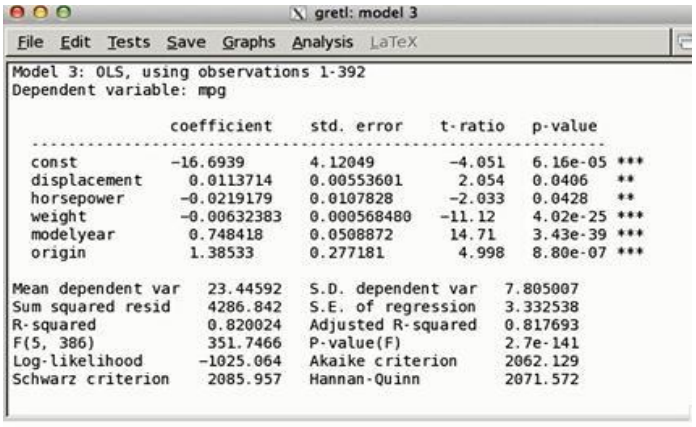
Réparation de R-quadratique

- Nous pouvons définir un nouveau

R-Quadratique qui seulement augmente quand la contribution est importante. Cette version s'appelle R-Quadratique ajusté

- Nous pouvons penser que c'est une fonction de R-Quadratique et du numéro de variables de la régression
- R-Quadratique ajusté =f(R-Quadratique,K)
- R-Quadratique ajusté seulement augmente quand la contribution de la variable est suffisamment grande.




Modèle final, Stratégie 2

- Le Modèle final est le modèle que maximise le R-Quadratique ajusté
- Ce modèle inclus cylinders, displacement, horsepower, weight, modelyear, origin

## VII.3 - Régression logistique

Variables dépendantes binaires

- Dans la session antérieure nous avons expliqué comment faire face à des variables quantitatives dépendantes.
- Mnt, analysons le cas ou la variable dépendante est binaire.
- Exemple : Acceptation d'un crédit (banque)

codification

- La codification des variables sera identique
- Affectons 1 et 0 pour la présence ou l'absence d'une condition.
- Y est une variable binaire

Exemple

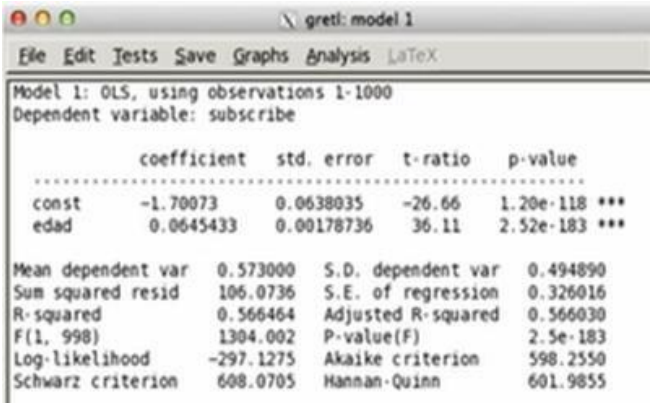
- Nous avons accès à un échantillon aléatoire de 1000 consommateurs dans une ville V.
- Imaginez que nous somme entrain d'étudier la décision d'inscription ou non à une revue.
- Nous voulons expliquer cette décision comme une fonction de l'Age du consommateur

Définitions

- Subscribe est la variable dépendante. Égale à 1 si le consommateur d'inscrit et 0 sinon.
- Age est la variable indépendante

Régression linaire

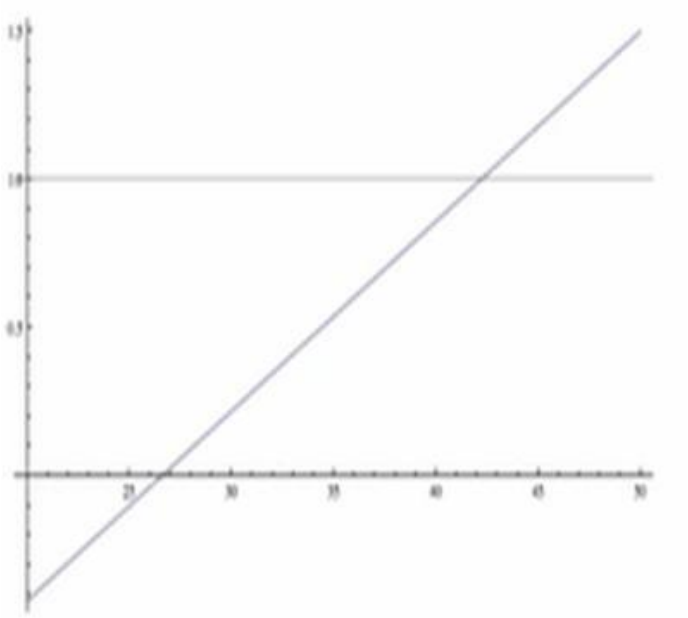
Subcribe=b+a*Age




- Subscribe=-1.70+0.064*Age
- Interprétation : p=-1.70+0.0064*Age (p la probabilité d'inscription)
- Interprétation de la pente est triviale : pour chaque année d'Age de plus la probabilité d'inscription augmente par 6,4%

problème

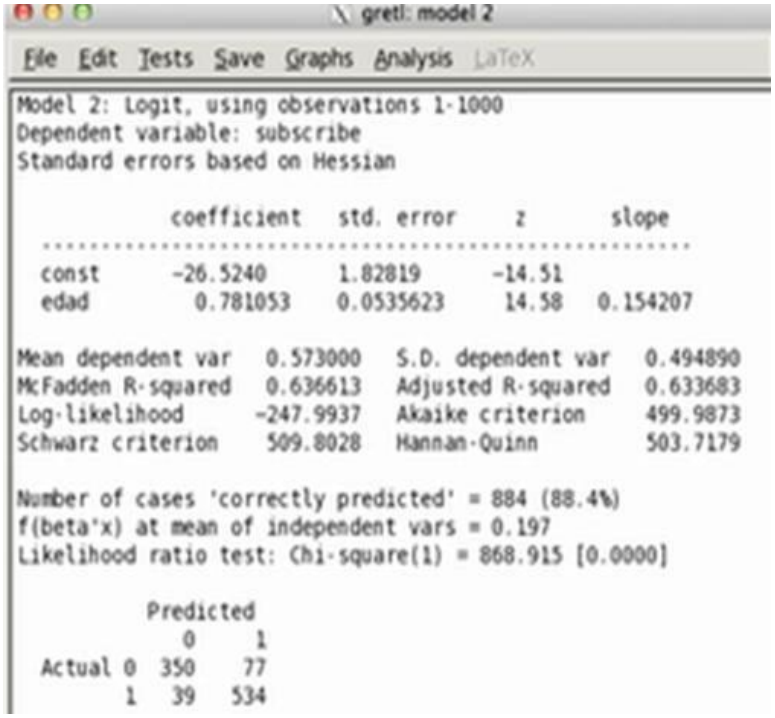
- Les problèmes cette équation surviennent lorsque nous essayons de faire des prévisions avec elle.
- La probabilité d'une personne de 35 ans s'inscrit dans la revue?
- P=-1.70+0.064*35=0.35 (pas de problème)
- Et de 25 ans? De 45 ans?
    - Vérifier que -0.10 et 1, 20



Solution

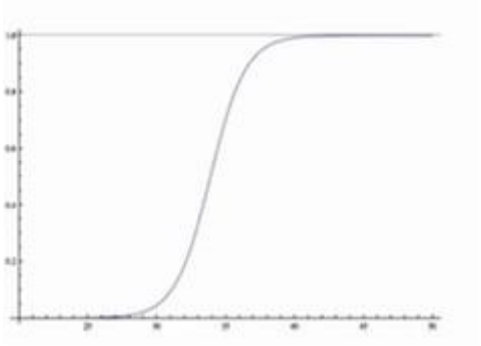
- Changer la spécification
    - Les valeurs de p doivent être dans [0,1]
    - P=f(Age)
    - Nous aurons besoin de deux choses:
        - f positive
        - f<=1
- f est une fonction non linéaire
- Positive peut être exponentielle
- P=exp(b+a*Age)
- F<=1 doit être p=exp(b+a*Age)/(1+exp(b+a*Age))
- Bye bye la linéarité
- Ln(p/(1-p))=b+a*Age

- Il se peut que la probabilité n'est pas une fonction linéaire de l'âge, mais une simple transformation de celui-ci
- C'est l'équation de la régression logistique




Régression logistique

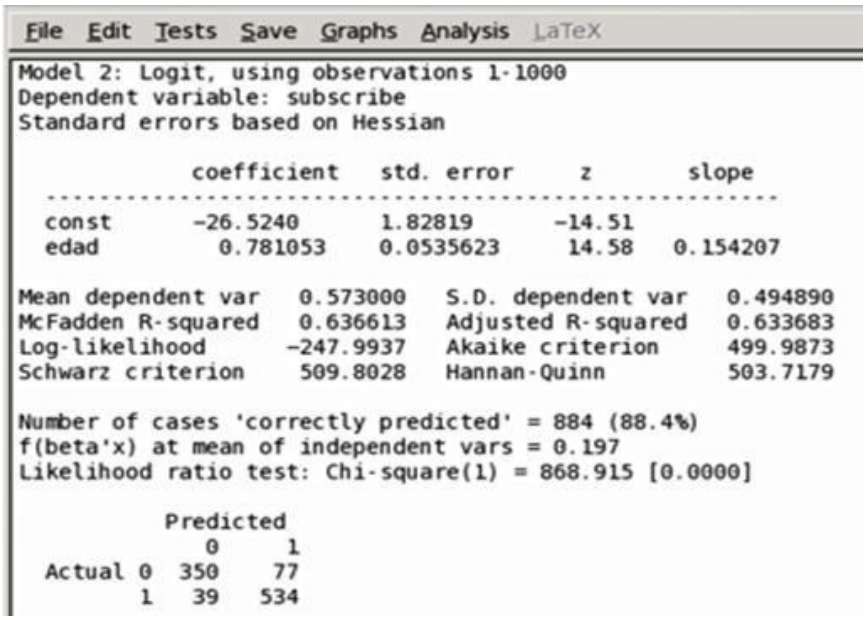
- L'équation est:
- Ln(p/(1-p))=-26.52+0.78*Age
- Exprimer en terme de probabilité
- p=exp(-26.52+0.78*Age)/((1+exp(26.52+0.78*Age)



Interprétation des coefficients et prévisions

- Nous avons expliqué pq nous n'avons pas utiliser la régression linéaire avec une variable dépendante est binaire
- Au lieu de régression linéaire on parle de régression logistique

- Ln(p/(1-p))=26.52+0.78*Age
- Si on pose y*=ln(p/(1-p))
- Y*=26.52+0.78*Age
- cela ressemble à une régression linéaire commune et courante.



File Edit Tests Save Graphs Analysis LaTeX

```
Model 2: Logit, using observations 1-1000
Dependent variable: subscribe
Standard errors based on Hessian

             coefficient   std. error     z       slope
  ---------------------------------------------------------
  const      -26.5240      1.82819      -14.51
  edad         0.781053    0.0535623     14.58   0.154207

Mean dependent var   0.573000   S.D. dependent var   0.494890
McFadden R-squared   0.636613   Adjusted R-squared   0.633683
Log-likelihood      -247.9937   Akaike criterion     499.9873
Schwarz criterion    509.8028   Hannan-Quinn         503.7179

Number of cases 'correctly predicted' = 884 (88.4%)
f(beta'x) at mean of independent vars = 0.197
Likelihood ratio test: Chi-square(1) = 868.915 [0.0000]

           Predicted
              0     1
  Actual 0  350    77
         1   39   534
```

- L'erreur standard =0.0535
- Dans l'intervalle de confiance de coefficient de l'age est 0.7810+/-2*0.0535
- Le 0 est exclus➔ statistiquement significative
- Nous pouvons vérifier le p-value (beaucoup de logiciel ne fournissent cette valeurs)

Quel changement?

- C'est quoi 0.78 dans l'équation?
- Ln(p/(1-p))=26.52+0.78*Age
- Pour chaque année additionnelle, Ln(p/(1-p)) augmente 0.78 unité …
- Mais c'est quoi Ln(p/(1-p)), pour cela utiliser Excel pour trouver p.

Régression logistique multiple

Le même principe que la régression linéaire multiple

## Références Bibliographiques

Jean-Michel Franco, Le Data Warehouse, le Data Mining, Eyrolles, 1996

Michael J.A. Berry et Gordon S. Linoff, Data Mining:Techniques appliquées au marketing, à la vente et aux services clients, Masson, 1997

René Lefébure et Gilles Venturi, Le data mining, Eyrolles, 1998

Pierre Lévine et Jean-Charles Pomerol, Systèmes interactifs d'aide à la décision et systèmes experts, Hermès, 1990

Jean-Charles Pomerol, Les systèmes experts, Hermès, 1988

Olivier Cérutti et Bruno Gattino, Indicateurs et tableaux de bord, Afnor, 1993

Hervé Sérieyx, Le big bang des organisations, Editions Calmann-Lévy, 1993