ARBRES DE DECISIÓ

Ruth Vilà Baños
Departament MIDE
ruth_vila@ub.edu

Universitat de Barcelona

Escola d'Hivern de Doctorat
Curs 2011-2012

Escola d' Hivern de Doctorat
Fac. Pedagogia
21, 22 i 23 de febrer de 2012

TEMES

1. Anàlisi multivariant de dades
2. Data Mining
3. Procés d'extracció del coneixement- KDD (Knowledge Discovery in Databases)
4. Arbres de decisió
5. El mètode de divisió
6. Tècniques on es poden utilitzar els arbres
7. Arbres de decisió amb SPSS
8. Posem-ho a la pràctica!

ANÀLISI MULTIVARIANT DE DADES

Hi ha variables explicatives i explicades?

- **NO**
- **SI**

- **Mètodes Descriptius**: *tècniques d'anàlisi de la interdependència*
- **Mètodes explicatius**: *tècniques d'anàlisi de la dependència*
- **Tècniques emergents**: mineria de dades

Pérez, c. (2004) *Técnicas de análisis multivariante de datos*. Madrid: Pearson educación

Mètodes Descriptius

- Mètodes multivariants de REDUCCIÓ DE LA DIMENSIÓ:
  - Variables quantitatives:
    - Components principals: *reduir la gran quantitat de variables en unes poques perfectament calculables.*
    - Anàlisi factorial: *reduir la gran quantitat de variables en variables fictícies, no observades ni mesurades.*
  - Variables qualitatives:
    - Anàlisi de correspondències múltiple: reduir en un mapa gràfic l'associació entre les categories.
- Mètodes multivariants de CLASSIFICACIÓ DE GRUPS:
  - Anàlisi de conglomerats (cluster): classificació automàtica de dades en grups homogenis no coneguts a priori.
- Escalament multidimensional: representació gràfica perceptual

Mètodes Explicatius

		VARIABLES INDEPENDENTS	
		Quantitatives	**Qualitatives**
VARIABLE DEPENDENT	**Quantitativa**	*Regressió lineal múltiple* *Anàlisi canònic:* correlació amb més d'una dependent	*ANOVA / MANOVA* *ANCOVA /MANCOVA* *Anàlisi conjunt:* relació amb la v.dependent ordinal
	Qualitativa	*Anàlisi discriminant:*predicció de la categoria en la que es situa *Models d'elecció discreta:* predicció de la probabilitat	

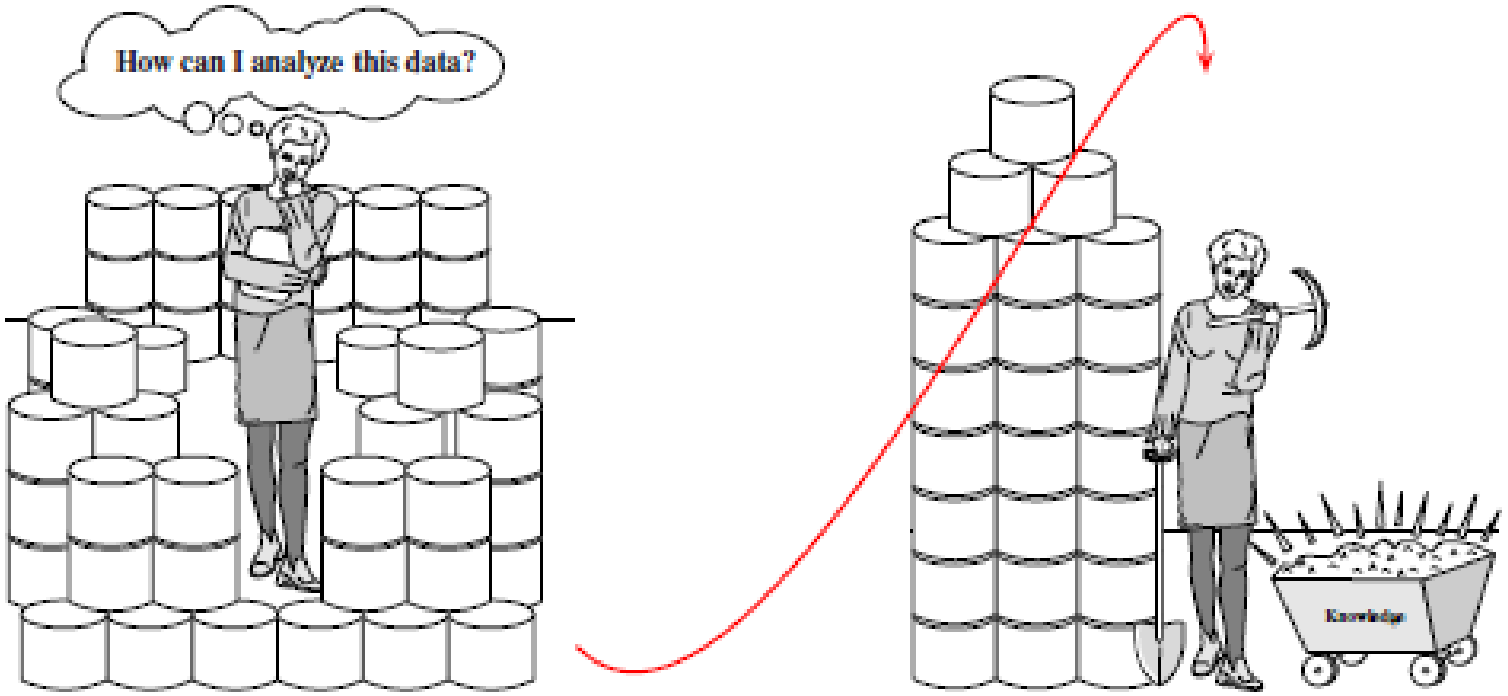
Tècniques Multivariants Emergents

- S'anomena Mineria de dades o DATA MINING
- Disponibilitat de grans volums de dades i eines informàtiques potents.
- Tècniques de data mining coincidents amb bona part de l'estadística multivariant:
  - Tècniques predictives: regressió, ANOVA, ANCOVA, mètodes bayesians, algoritmes genètics, arbres de decisió, xarxes neuronals, ...
  - Tècniques descriptives: clusters, segmentació, escalament multidimensional, ...

Pérez, C. i Santín, D. (2008) *Minería de datos. Técnicas y herramientas*. Madrid: Thomson

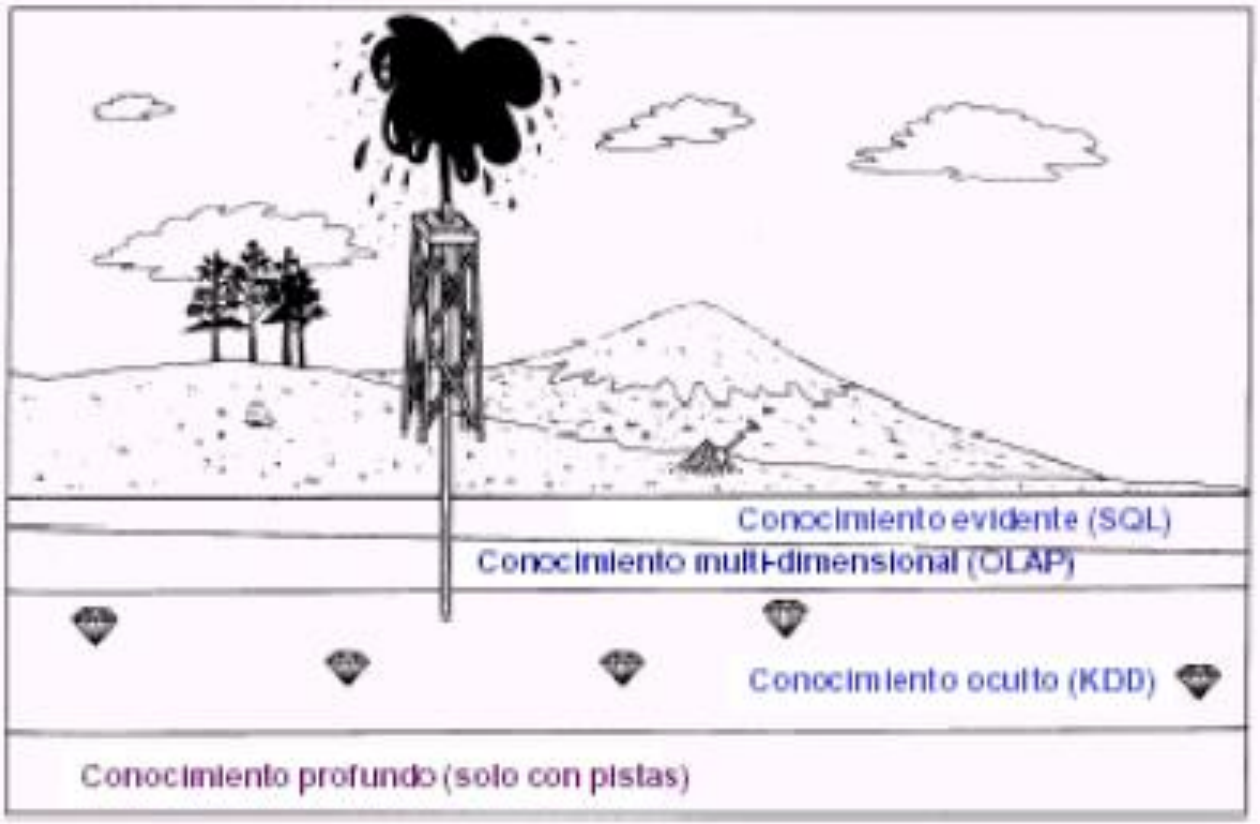
DATA MINING

Procés de descobriment de noves i significatives relacions, patrons i tendències en examinar grans quantitats de dades.



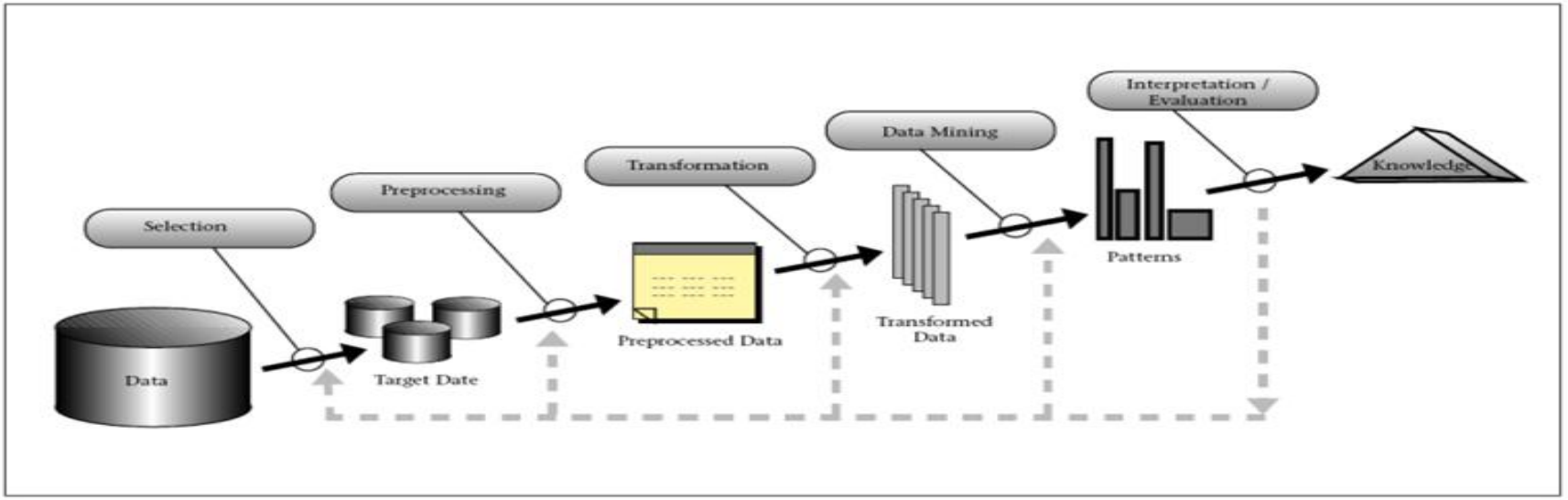
Han, J., Kamber, M i Pei, J. (2012) *Data mining. Concepts and techniques*. New York: Morgan Kaufmann

PROCÉS D'EXTRACCIÓ DEL CONEIXEMENT(KDD)



SQL: Structured Query Language
OLAP: Online Analytical Processing
KDD: Knowledge Discovery on Databases

Knowedge Discovery in Databases



1. Fase de selecció
2. Fase d'exploració
3. Fase de neteja i transformació de dades
4. Fase de mineria de dades
5. Fase de validació

PREPARACIÓ DE LES DADES

MINERIA DE DADES

Selecció

- Integració i recopilació de les dades
- Determinar les fonts d'informació
- Identificació i selecció de les variables rellevants
- Aplicació tècniques de mostratge

Exploració

- Comprovació dels supòsits dels mètodes multivariants per la mineria de dades:
  - Normalitat
  - Homocedasticitat
  - Variables linealment independents (correlació)
  - Linealitat
- Anàlisi exploratòria de dades:
  - Exploració visual: histogrames, diagrama de tija i fulles, gràfic de caixes i bigotis, de dispersió, ...
  - Exploració formal:
    - Simetria i normalitat
    - Correlacions entre variables

Neteja i transformació de dades

- Valors atípics (outliers), valors que falten (missing), errades, ...
- Anàlisi de la influència d'aquests valors atípics
- S'eliminen o es corregeixen les errades.

- Si és necessari es fa la transformació d'algunes dades.

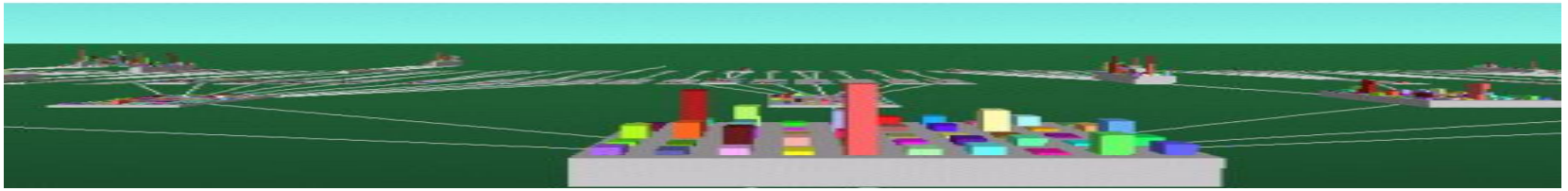
Tècnica de mineria de dades

- Decidir quina tasca es pretén: classificar, agrupar, ...
- Elecció de la tècnica:
  - Descriptiva: totes les variables tenen el mateix estatus.
  - Predictiva: es pot diferenciar entre variables dependent i independents, partint d'un coneixement teòric previ.

- Els arbres de decisió són predictius i de classificació: *tècniques de classificació ad hoc*, classifiquen individus o observacions dintre de grups prèviament definits.

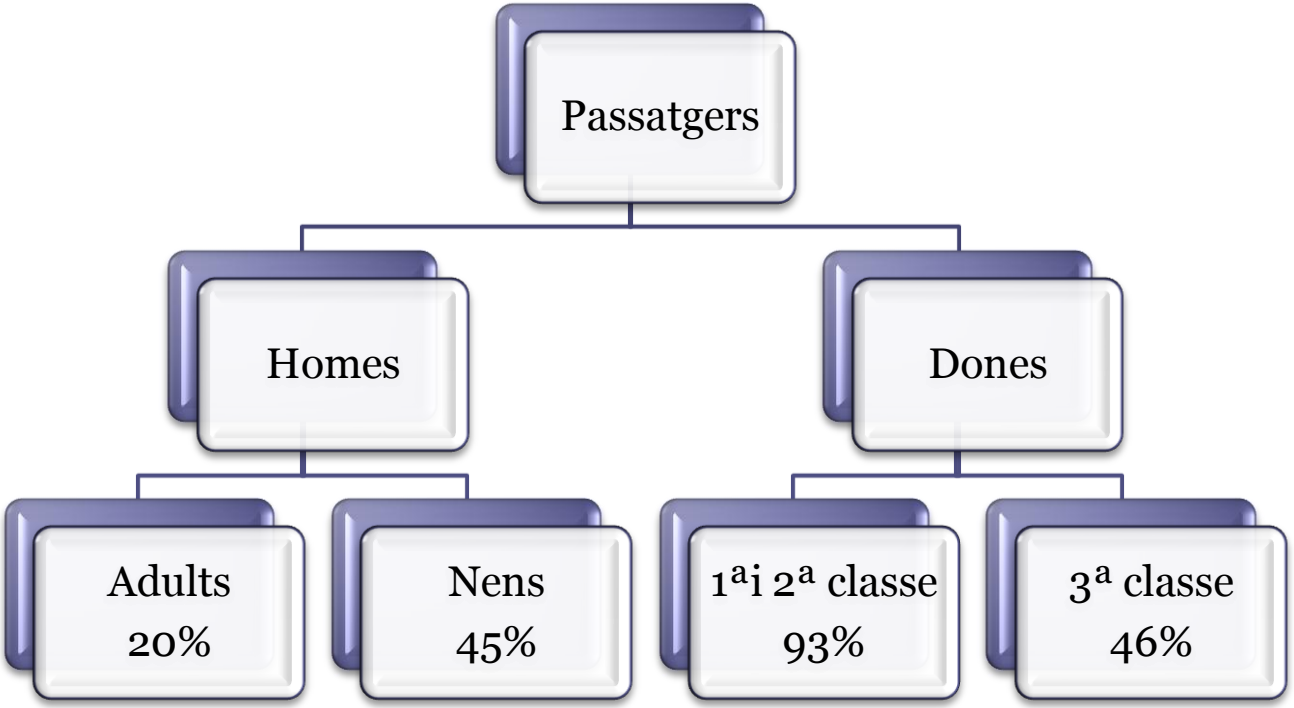
ARBRES DE DECISIÓ

- Arbres de classificació o arbres de regressió.
- Grans mostres, per revelar formes complexes en l'estructura que no es detecten amb mètodes convencionals de regressió.
- La variable dependent i les independents poden ser nominals, ordinals o d'escala.
- Fàcils d'interpretar.
- Procés de segmentació en funció de la divisió més discriminant dels criteris establerts.



Un exemple...

VD = grau de supervivència al Titanic



Pérez, C. i Santín, D. (2008) *Minería de datos. Técnicas y herramientas*. Madrid: Thomson

EL MÈTODE DE DIVISIÓ

- Arbres CHAID
- Arbres CRT
- Arbres QUEST

Arbres CHAID (*Chi-square automatic interaction detector*)

- Mètode exploratori d'anàlisi de dades per identificar variables importants.
- Finalitat de segmentació, anàlisi descriptiu o previ a altres anàlisis.
- La variable dependent pot ser qualitativa o quantitativa.
- Chaid exhaustiu tracta totes les variables per igual (independentment del tipus i nºcategories)
- Pot produir divisions de més de dos grups.

Arbres CRT *(Classification and regression trees)*

- Alternativa al chaid exhaustiu , superant algunes limitacions de la versió inicial.
- Apropiat per arbres de classificació (VD qualitativa) o de regressió (VD quantitativa).
- Els arbres són binaris.
- Permet elegir entre diferents coeficients i mesures, superant alguns “falsos positius” de la X^2.

Arbres QUEST *(Quick, unbiased, efficient, statistical tree)*

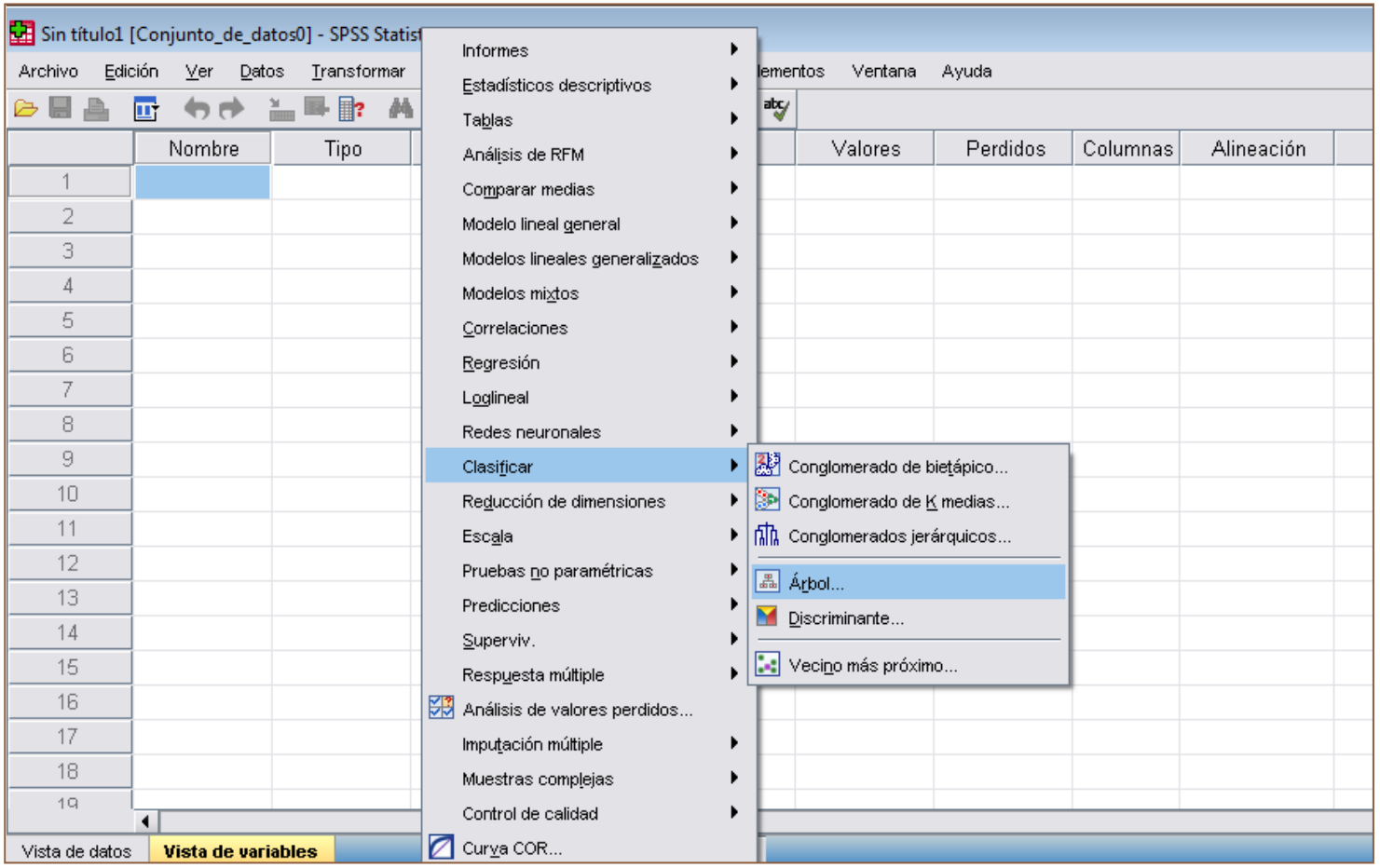
- Algoritme creat per superar dues limitacions dels anteriors:
  - Complexitat computacional
  - Biaixos en la selecció de variables: tendència a seleccionar aquelles que tenen un major nombre de categories.
- Arbre binari.

TÈCNIQUES ON ES PODEN UTILITZAR ELS ARBRES

- SEGMENTACIÓ: identifica individus en un grup específic.
- ESTRATIFICACIÓ: assigna casos a una categoria (alt, mig, o baix risc)
- PREDICCIÓ: crea regles i les utilitza per predir el futur.
- REDUCCIÓ DE DADES I CLASSIFICACIÓ DE VARIABLES: selecció de variables predictores
- IDENTIFICACIÓ D'INTERACCIÓ: relacions en subgrups específics
- FUSIÓ DE CATEGORIES I DISCRETITZACIÓ DE V. CONTÍNUES: recodifica perdent poca informació

Pérez, C. i Santín, D. (2008) *Minería de datos. Técnicas y herramientas*. Madrid: Thomson

ARBRES DE DECISIÓ AMB SPSS

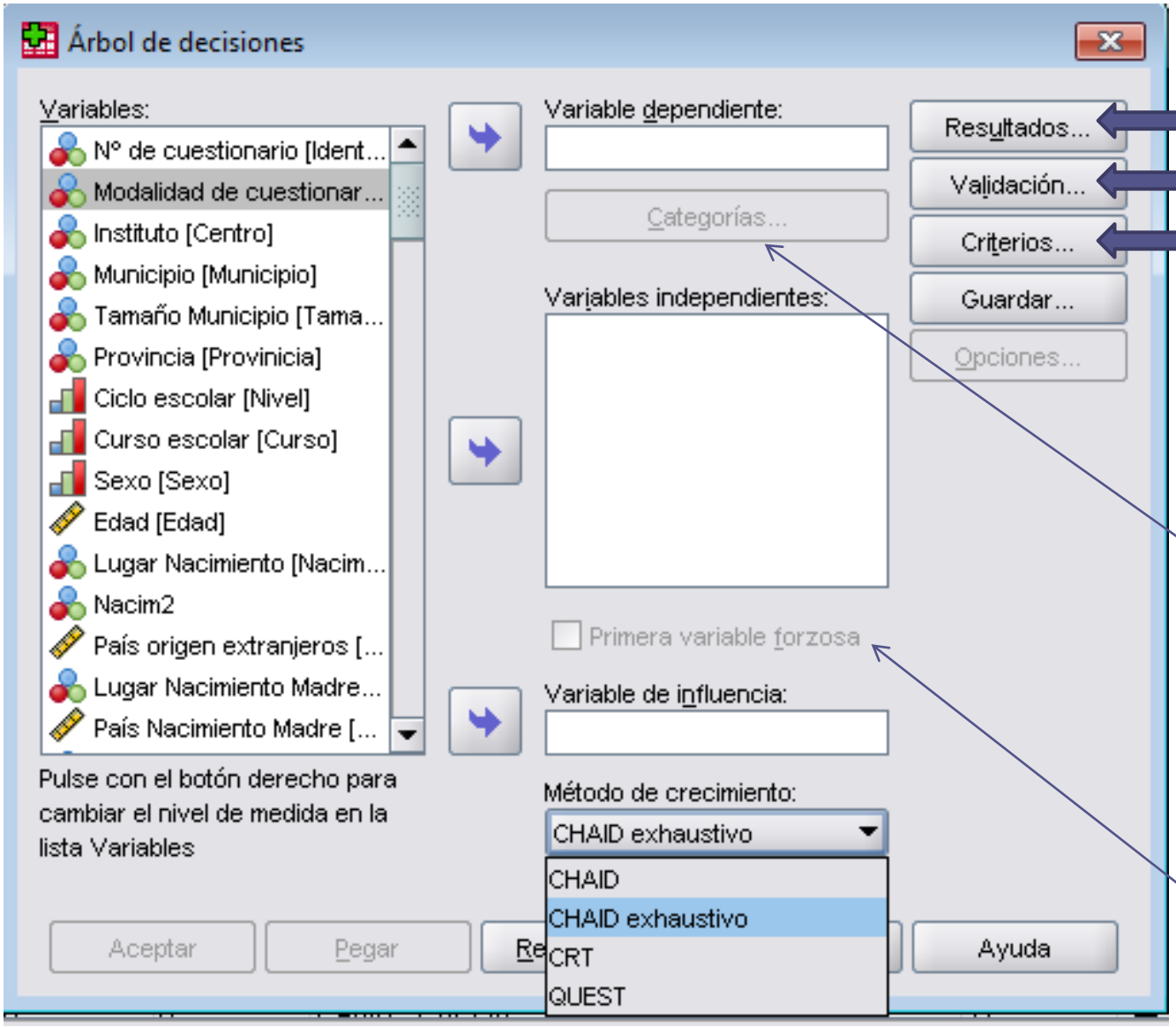


Atenció!

Reviseu les escales de mesura assignades a la matriu de dades d'SPSS, pot afectar a l'arbre.

Seleccionem 1 variable dependent i una o més d'independents.

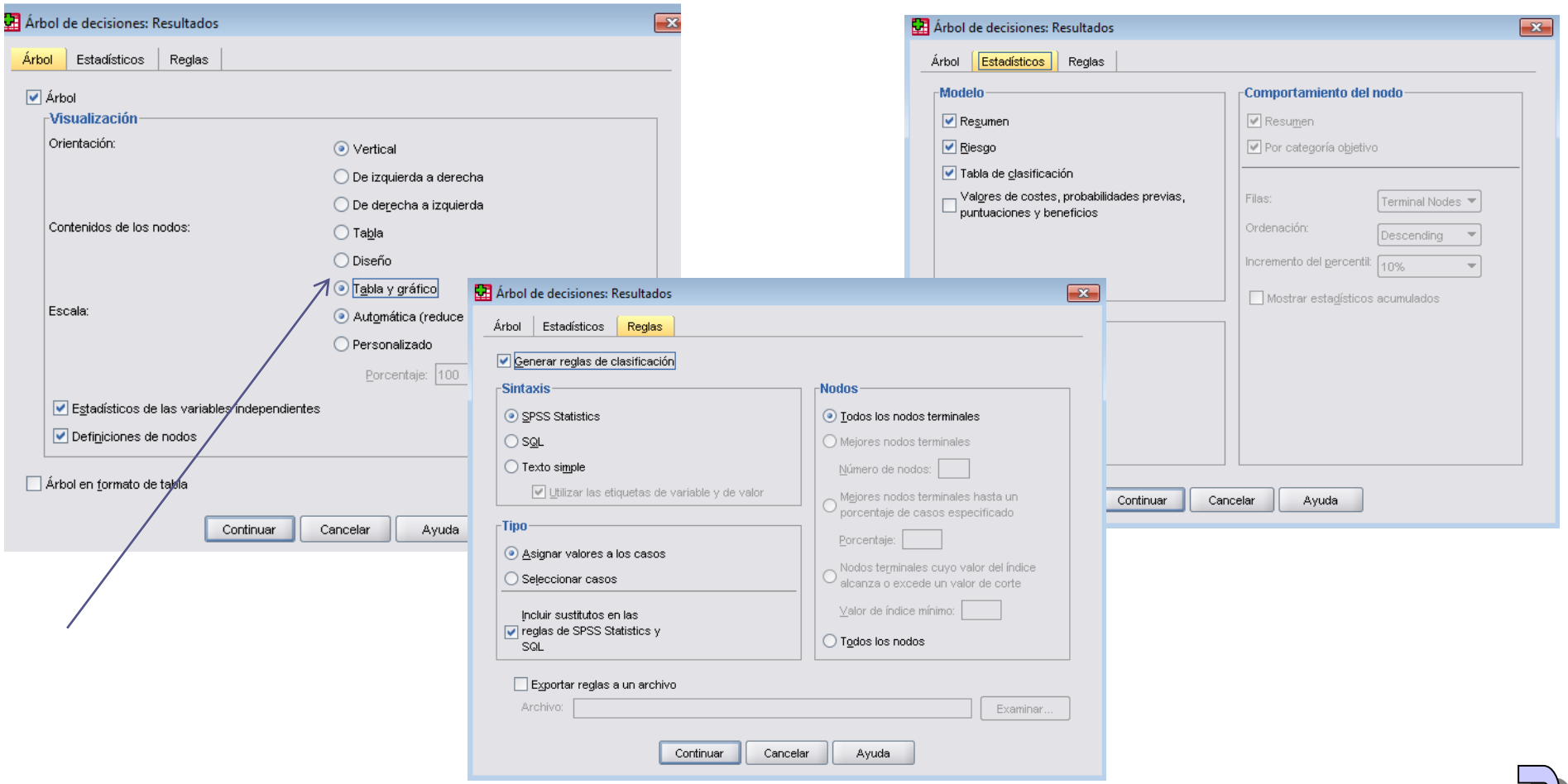
Seleccionem el mètode de creixement: CHAID, CRT O QUEST



Per seleccionar una o més categories d'interès

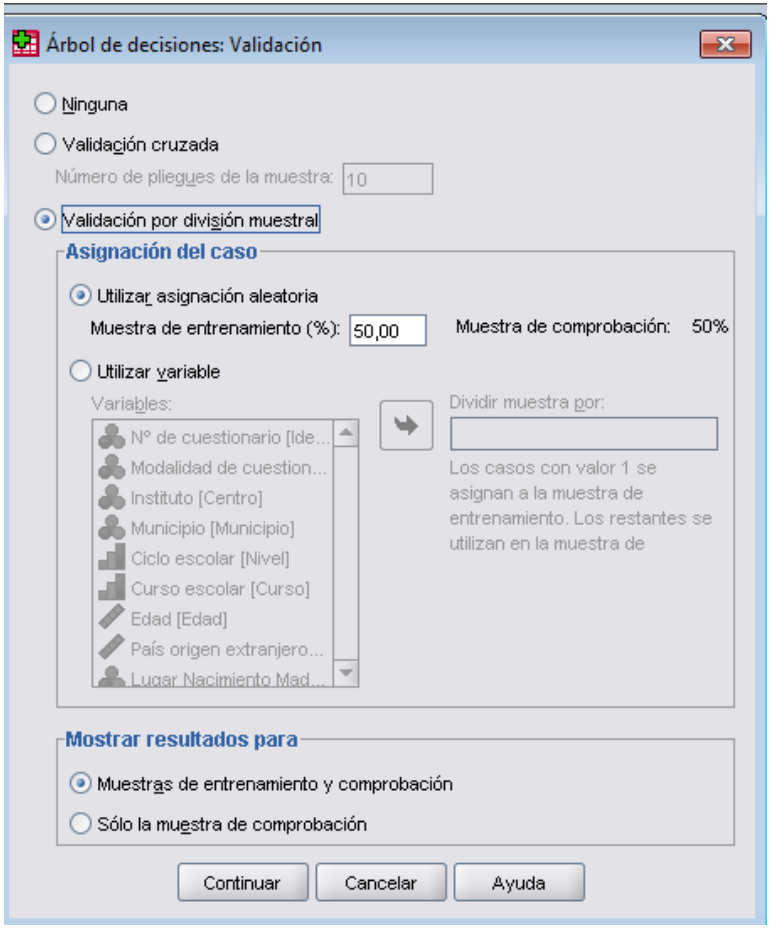
Per forçar que la primera independent de la llista sigui la primera v de divisió en l'arbre

Forma, estadístics, gràfics i regles



Validació de l'arbre

- Per avaluar la bondat de l'estructura d'arbre en generalitzar a la població.
- 2 mètodes:
  - Validació creuada: genera submostres i es fa el promig de risc entre totes elles (*pliegues*).
  - Validació per divisió mostral: fa una mostra d'*entrenament* i posa a prova el model amb una mostra de *comprovació*.

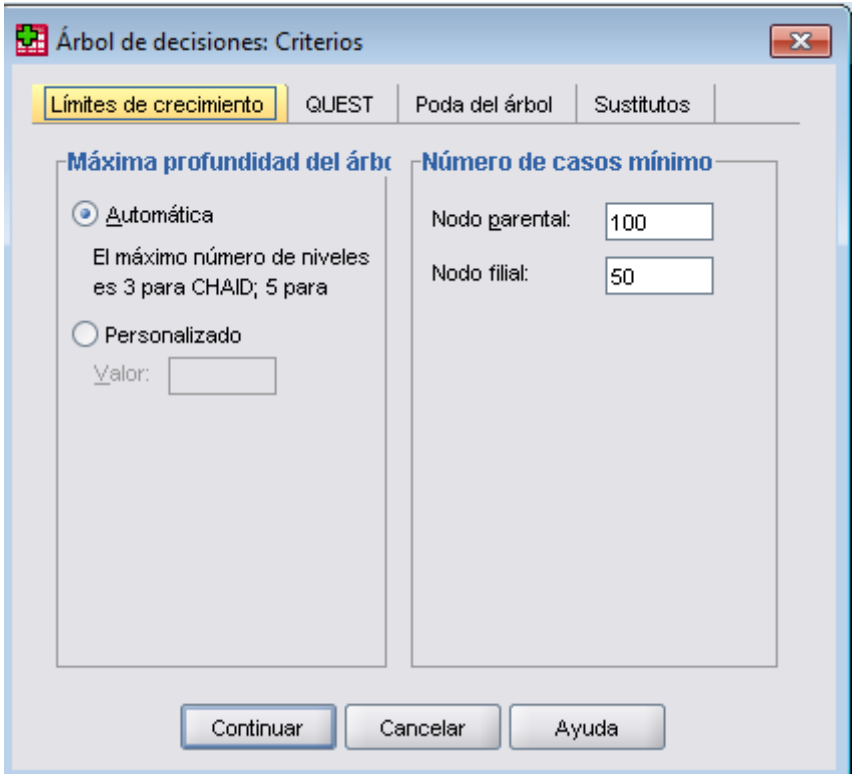


Criteris de creixement de l'arbre

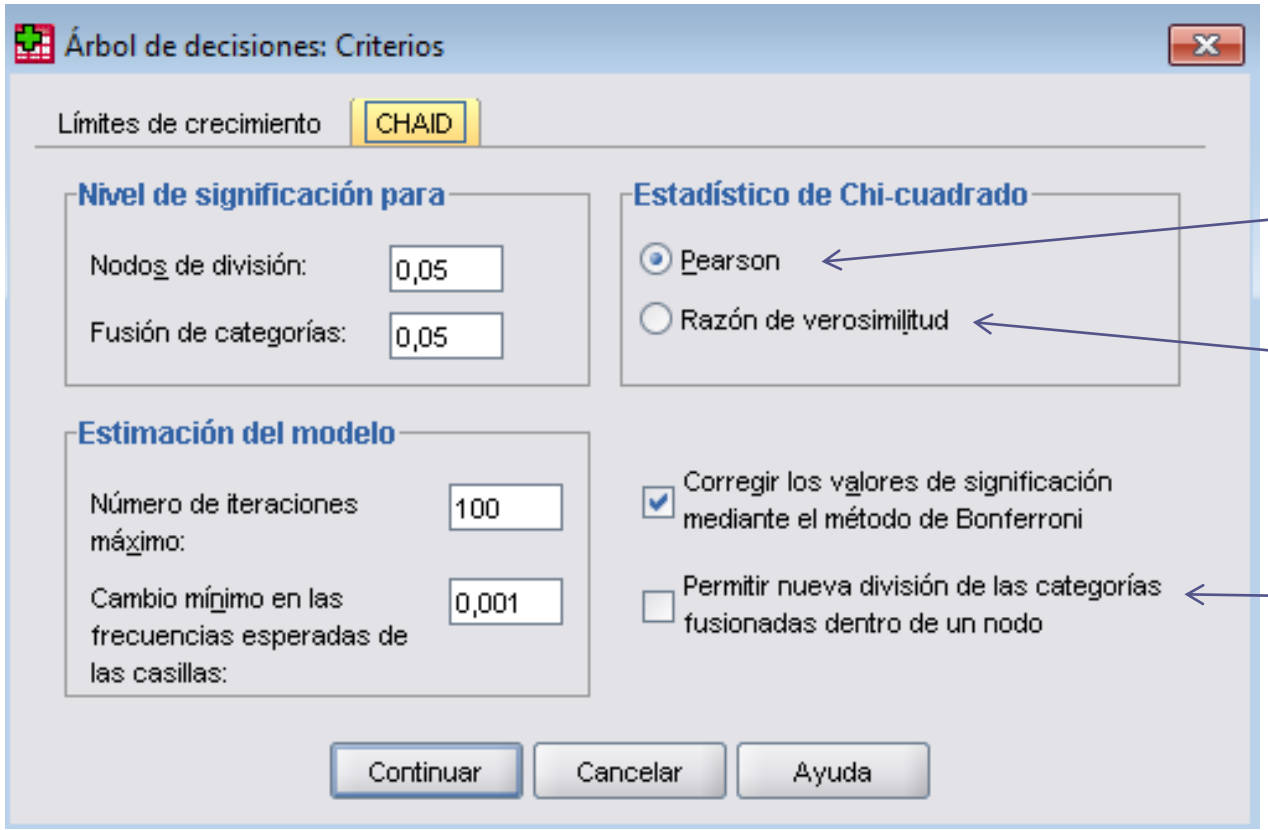
Nivells de l'arbre i nombre de casos mínim pels nodes.

Segons el mètode de divisió seleccionat:

- CHAID
- CRT
- QUEST



Criteris de creixement per CHAID

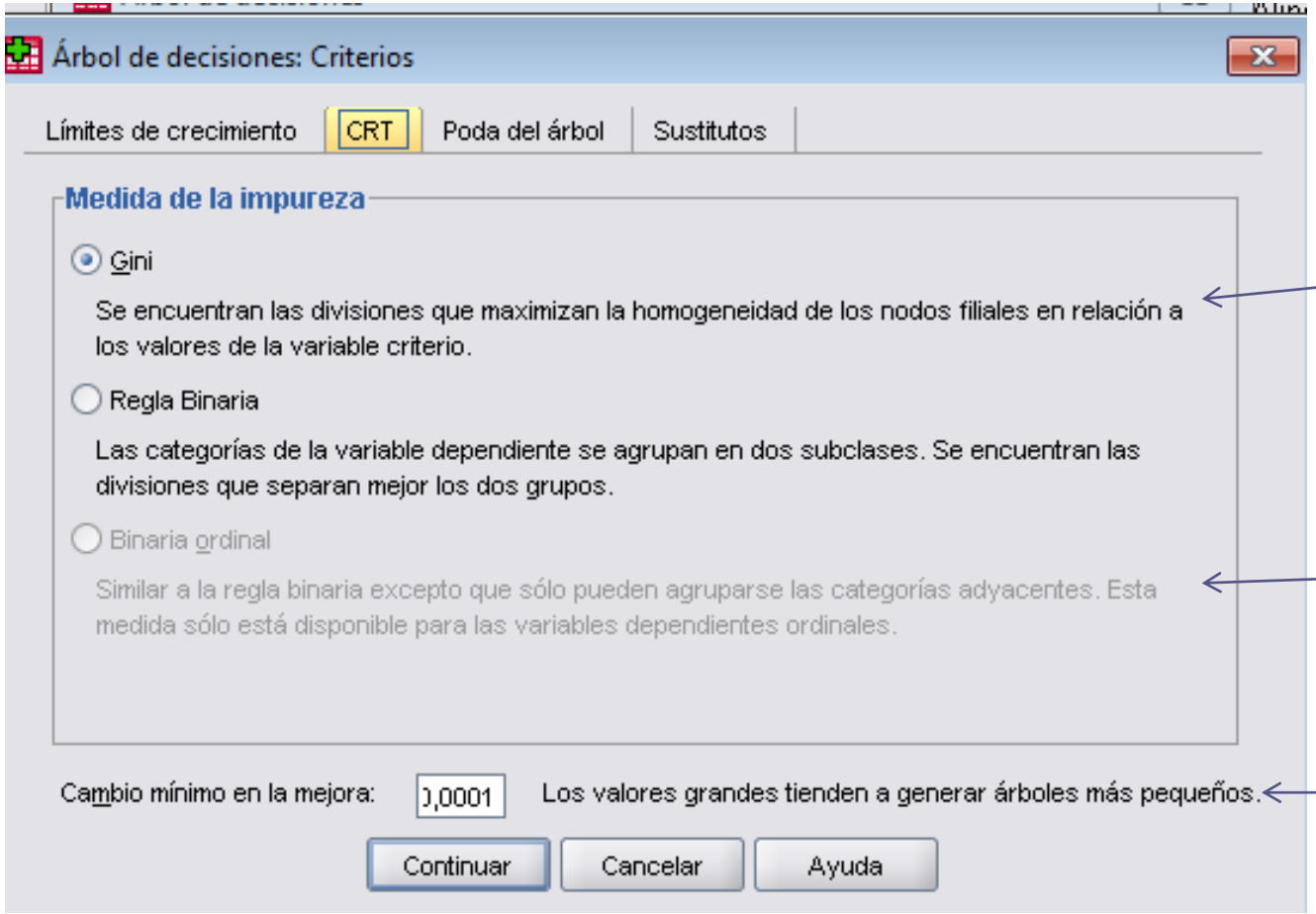


Càlculs ràpids i mostres grans

Robust o mostres petites

Permet simplificar l'arbre

Criteris de creixement per CRT



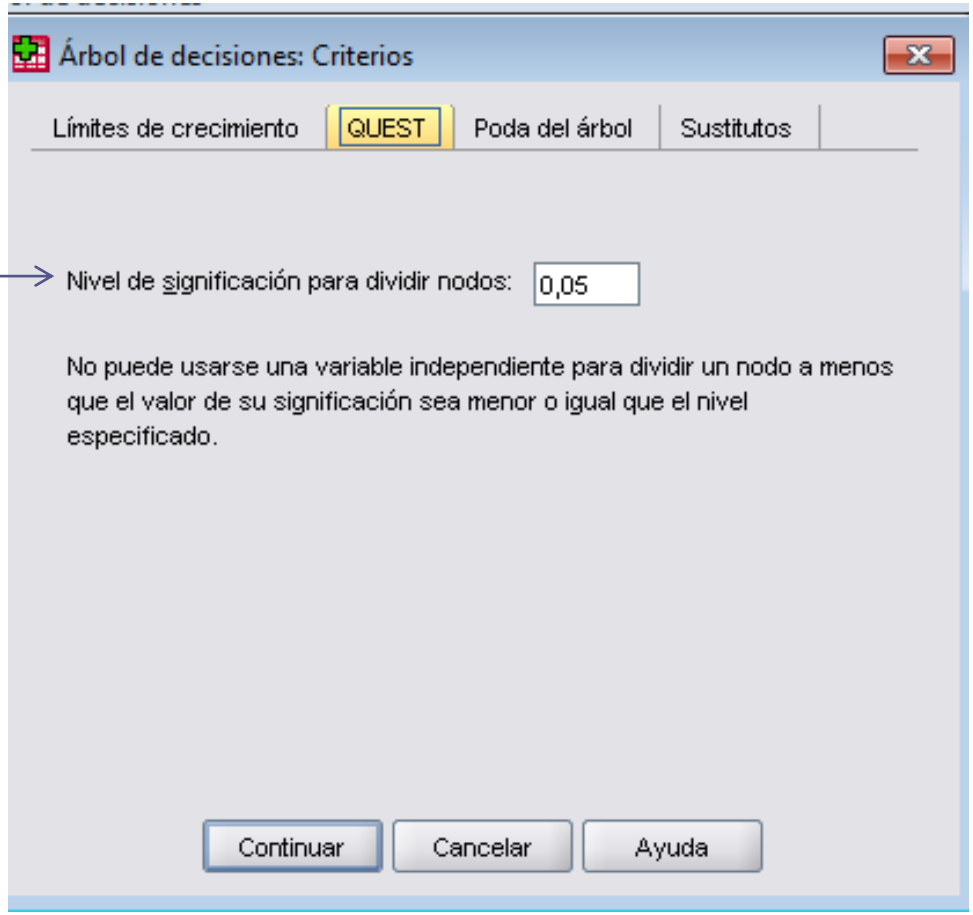
Per variables categòriques

Només per a ordinals

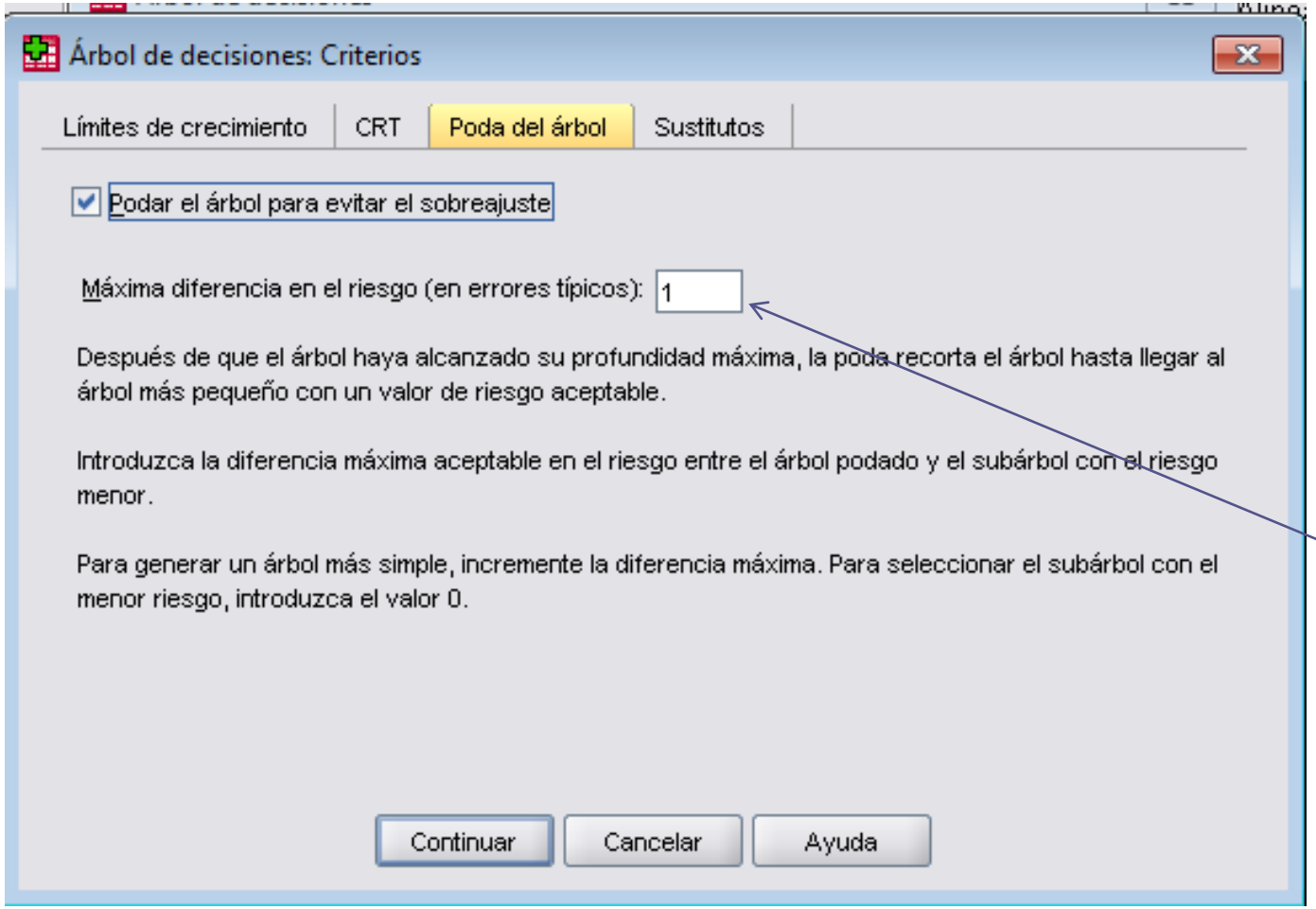
Reducció mínima de la impuresa per dividir els nodes

Criteris de creixement per QUEST

Nivell de significació (0-1).
A menor significació, es tendirà a excloure més variables independents del model final.



La “poda” de l’arbre. CRT i QUEST



Retall de l’arbre *(poda)* automàtica per obtenir un subarbre més petit amb el risc especificat

Mínim risc = 0

Moltes gràcies!