TFC MINERIA DE DADES
26/07/2005

Introducció d'una empresa a l'extracció del coneixement a partir d'unes dades.

Alumne: Cristina Collell Ventepani
ETIG
Consultor: Ramon Caihuelas Quiles
10-01-2005

INDEX

PART PRÀCTICA

0. TEMPORITZACIÓ I OBJECTIUS

TFC àrea Mineria de Dades: Implantació d'un projecte de Knowledge Center amb una eina comercial.

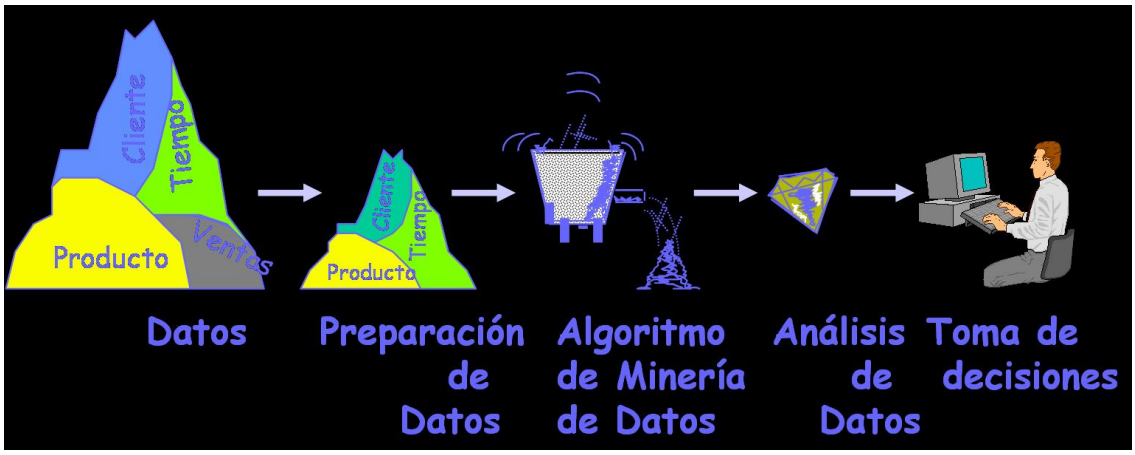
Situació (real):
L'editorial Motorpress Ibérica, amb molta tradició en revistes del sector del motor, ha comprat una revista, MaxiTuning, dedicada a cotxes preparats.
De curta vida, 6 anys, aquesta revista ha triomfat entre els lectors per si mateixa, sense haver fet cap estudi, ni sorgir cap preocupació d'estudiar el públic a la que va dirigida ni qui la llegeix. Encara així, amb molts lectors i compradors, s'ha situat en la primera revista més venuda del motor i la sexta si contem totes les àrees.
Però ara hi han moltes editorials que s'han apuntat a la moda tuning, i estan sortint revistes com bolets.
Situació (hipotètica):
Els directius demanen dos coses:
1. Que els formem en que és realment el Data Mining, bé, en general que formem als directius (principalment de marketing i informàtica) en el que hagin de saber sobre el descobriment del coneixement en les bases de dades, que puguin tenir una bona font d'informació per a poder entendre del tema.
2. Que a més, introduint-los a l'hora àmpliament en el potencial del paquet Synera, aprofitem les dades (reals) que l'editorial ha extret de la seva pagina web i de terminals disposats a la última macroconcentració realitzada el passat Setembre on es desplacen fanàtics de tota Espanya, sobre els seguidors del tuning per a poder fer les primeres campanyes dirigides als potencials lectors i començar així a lluitar contra la competència. A més d'una enquesta realitzada des de la pròpia revista sobre les que també llegeixen a més de la pròpia Maxi Tuning. Aquestes dades, encara que escasses ja que no cobreixen extensament les característiques i aficions que pot tenir una persona, serviran per a fer un primer aproximament amb el programari que es vol introduir a l'empresa: el paquet Synera.

Les dades reals de que disposo son:

- *3.219 entrades que donen edat-sexe-preguntes bàsiques sobre gustos-i revistes llegides (dades recollides des d'un formulari de la revista).*
- *7.862 entrades que responen varies preguntes sobre costums i característiques varies sobre els potencials "tuneros" (dades recollides des d' Internet i des de les terminals de la concentració de vehicles –visitants i participants-).*

Tasca	Duració	Inici	Fi
Inici. Elaboració Pla de Treball	7 dies	16-09-04	22-09-04
Repassar documentació DataMining	*1 dia*	*16-09-04*	*16-09-04*
Llegir documentació tauler	*1 dia*	*17-09-04*	*17-09-04*
Búsqueda paraules clau Internet	*1 dia*	*18-09-04*	*18-09-04*

Decisió tema	*1 dia*	*19-09-04*	*19-09-04*
Elaboració pla de treball	*1 dia*	*20-09-04*	*20-09-04*
Instal·lar Synera	*1 dia*	*21-09-04*	*21-09-04*
Repassar Pac1. Entrega.	*1 dia*	*22-09-04*	*22-09-04*
PAC 1: PLA DE TREBALL	0 dies	22-09-04	22-09-04
Documentació teòrica dels aspectes que impliquen la gestió del coneixement.	22 dies	23-09-04	14-10-04
L'estat de les tecnologies de Transformació de dades i la seva utilitat i el seu entorn (data Warehouse, OLAP, OLTP, Data Mining...)	*6 dies*	*23-09-04*	*28-09-04*
Processos en la cerca de coneixement	*3 dies*	*29-09-04*	*01-10-04*
Cicle de vida	*3 dies*	*02-10-04*	*04-10-04*
Preparació de dades i tasques prèvies a l'anàlisi. Models	*3 dies*	*05-10-04*	*07-10-04*
Tècniques d'estudi dels resultats i tipus de mètodes per a elaborar-los	*7 dies*	*08-10-04*	*14-10-04*
Documentació teòrica sobre Synera	14 dies	15-10-04	30-10-04
Què es Synera.	*2 dies*	*15-10-04*	*16-10-04*
El perquè del Synera	*2 dies*	*17-10-04*	*18-10-04*
Ús del Synera actualment arreu	*3 dies*	*19-10-04*	*21-10-04*
Preparació docs sobre funcionament (per a l'empresa i repàs propi)	*9 dies*	*22-10-04*	*30-10-04*
Creació document Pac 2. Repàs i entrega.	3 dies	31-10-04	02-11-04
PAC 2: FORMACIO EMPRESA DATAWAREHOUSE & SYNERA	0 dies	02-11-04	02-11-04
Plantejament problema	6 dies	03-11-04	08-11-04
Introducció dades de paper a editor	*6 dies*	*03-11-04*	*08-11-04*
Preparació dades	10 dies	09-11-04	18-04-04
Extracció i transformació	*4 dies*	*09-11-04*	*12-11-04*
Neteja i anàlisis	*4 dies*	*13-11-04*	*16-11-04*
Carrega	*2 dies*	*17-11-04*	*18-11-04*
Creació Resultats	18 dies	19-11-04	06-12-04
Models	*12 dies*	*19-11-04*	*30-12-04*
Anàlisi resultats. Revisió.	*4 dies*	*01-12-04*	*04-12-04*
Valoració final	*2 dies*	*05-12-04*	*06-12-04*
Creació document Pac 3. Repàs i entrega	3 dies	07-12-04	09-12-04
PAC 3: TREBALL DADES EMPRESA	0 dies	09-12-04	09-12-04
Preparació memòria.	12 dies	10-12-04	21-12-04
Preparació presentació.	12 dies	22-12-04	02-01-04
Revisió final i correccions.	8 dies	03-01-04	10-01-04
ENTREGA FINAL	0 dies	10-01-05	10-01-05



"Como ocurre en algunas fases del KDD, en una mina se desechan enormes cantidades de material inservible antes de que oro o diamantes sean encontrados."

1. INTRODUCCIÓ

En un mercat tan competitiu i canviant com l'actual es molt important usar tots els recursos per a extreure informació i coneixement de les dades de les que disposem.
De prendre decisions importants depèn molt l'èxit de les empreses. I extreure coneixement de les dades que genera contínuament una empresa es decisiu per a tenir prendre bones decisions.

I no es només el tenir accés de manera ràpida i flexible a les dades més rellevants.

La gestió del coneixement cobreix els aspectes de us de la informació per a la presa de decisions, des de la seva extracció en els sistemes, depuració, transformació, el disseny de estructures de dades per a l'emmagatzemament de dades fins a l'explotació de la informació mitjançant eines comercials de fàcil us per a usuaris que no tenen perquè dominar aquest àmbit.

La Mineria de Dades està dins d'aquesta visió d'àrees d'estudis dirigits al suport de funcions analítiques de la gestió empresarial per a la planificació estratègica i tàctica i recolzament per a la presa de decisions.
I aquí entra el Data Warehousing, que encara que no es una manera de emmagatzemar les dades completament imprescindible, es molt útil per a emmagatzemar les dades de manera que es puguin analitzar més eficientment i usar tecnologies analítiques especialitzades. Dit al revés, la Mineria de Dades és una completa manera de extreure informació de les dades emmagatzemades en un DW.

Però de totes aquestes paraules, i de lo més important, de com extreure coneixement, parlarem en aquest treball.

2. GESTIO DEL CONEIXEMENT

"Gestió del Coneixement és el procés d'administrar contínuament coneixement de tot tipus per a satisfer les necessitats presents o futures, per a identificar i explotar recursos de coneixement tant existents com adquirits i per a desenvolupar noves oportunitats."

El coneixement s'ha convertit en un factor emergent i diferenciadors entre la pobresa i la riquesa, que ha dut amb si una nova disciplina per a la seva administració, distribució i ús: la Gestió del Coneixement.

La Gestió del Coneixement és una corrent modelitzadora de la transformació de les organitzacions introduint la consideració de un altre recurs més (el coneixement) per a donar resposta a les noves demandes de canvi i millora, i per a aconseguir mantenir posicions competitives usant de manera intensiva les capacitats de les persones i de les tecnologies de la informació.

Algunes característiques del Coneixement:

- El coneixement es una capacitat humana. La seva transmissió implica un procés intel·lectual d'ensenyança i aprenentatge. Transmetre informació és molt més fàcil que transmetre coneixement. Això implica que quan parlem de gestió del coneixement volem dir que ajudem a aquestes persones a fer-ho.
- El coneixement no té valor si roman estàtic. Només genera valor en la mesura en que es transmeteix o transforma.
- El coneixement genera coneixement mitjançant l'ús de la capacitat de raonament.
- El coneixement té una estructura i es elaborat, implica l'existència de xarxes de relacions semàntiques entre entitats abstractes o materials.
- El coneixement es sempre esclau d'un context. Per tant al transmetre's és necessari que el emissor conegui el context o model del món del receptor.
- El coneixement por ser explícit (es pot recollir, manipular i transferir amb facilitat) o tàcit (p.ex. l'experiència acumulada de sentiments).
- El coneixement por estar formalitzat en diversos graus, poden ser també informal (com p. ex. la major part del coneixement transferit verbalment.

I aquest coneixement ha de expandir-se, les organitzacions han de crear una cultura organitzacional que faciliti que es comparteixi el coneixement i veure que les tecnologies son la millor manera per a això.

La Gestió per Coneixement es una altra manera de respondre al mateix fenomen, però donant importància al valor del coneixement com element estratègic que condiciona i configura l'organització i el seu modela la gestió i el desenvolupament de l'empresa, els seus productes i serveis, i la xarxa dels seus col·laboradors, com criteris clau en la missió i visió del valor del negoci, en la contribució de les persones i les seves responsabilitats, en l'organització dels equips de treball, i en el desenvolupament de l'estratègia orientada a explorar i explotar el coneixement.

La Gestió per Coneixement pretén configurar el desenvolupament i explotació més dinàmica, intel·ligent i eficaç dels recursos humans, com agents únics operadors del coneixement, constitueixen el substrat fonamental del desenvolupament organitzacional intel·ligent.
Les tecnologies de la informació son les considerades els mitjans per a l'explotació possible del coneixement.

Les condicions necessàries per a la creació d'un entorn de coneixement com una xarxa superior que enllaça els recursos constituïts per:

- La qualitat del recurs humà.
- **La capacitat de gestionar informació.**
- L'habilitat del model organitzatiu per implementar e integrar les eines, tècniques i mètodes adequats.

De la capacitat de gestionar i extreure la informació dependrà en bona manera que puguem usar-la per a la nostra finalitat: extraure coneixement. El coneixement es construeix a partir de la informació rebuda, s'emmagatzema en contenidors de informació i es transmet també a través de missatges amb contingut informatiu.
Les altre dos condicions serien objecte d'un altre treball.

2.1. Processos de la Gestió del Coneixement:

Transformació de la Informació en Coneixement (procés continu):

1. Generació de coneixement: crear o desenvolupar un coneixement necessari que fins al moment no es té.
2. Captura/Adquisició: importar i recol·lectar la informació. Es poden usar per a això elements típicament humans o automàtics com les bases de dades.
3. Organització: filtrar la informació, reconèixer lo que es important i lo que no, analitzar i validar-la. També es poden usar elements humans i automàtics.
4. Búsqueda i utilització: un cop seleccionada, organitzada, categoritzada i relacionada de la informació, hem de posar-la a disposició de qui la necessiti.
5. Publicació.
6. Distribució. La informació, ja convertida en material, al interactuar amb la persona li permetrà crear coneixement.

3. DATA WAREHOUSE

3.1 INTRODUCCIO AL CONCEPTE DATA WAREHOUSE (o dipòsit de dades)

W.H. Inmon: *" Data Warehouse es un sistema orientado al usuario final, integrado, con variaciones de tiempo y sobre todo una colección de datos como soporte al proceso de toma de decisiones".*

Bàsicament el procés que es segueix es que, a partir de les dades històriques es construeix una plataforma solida (organitzant i emmagatzemant les dades) per a fer l'anàlisi (mitjançant un procés analític informàtic).
Al fer així el DW queda com una col·lecció de dades orientat a temes, integrat, no volàtil, de temps variant, que s'usa per al suport del procés de presa de decisions gerencials.
Les diferencies entre les dades operacionals (els extrets directament de les aplicacions de producció) de les dades de negoci emmagatzemades en un DW són:

Base de Datos Operacional	Data Warehouse
Datos Operacionales	Datos del negocio para Información
Orientado a la aplicación	Orientado al sujeto
Actual	Actual + histórico
Detallada	Detallada + más resumida
Cambia continuamente	Estable

Hi ha que deixar clar que les dades del DW son quasi sempre dades trobades de les aplicacions operacionals.

A més hi ha que tenir en compte al crear una DW les grans diferencies amb les Bases de Dades operacionals de com els usuaris en fan us, ja que serà primordial a l'hora d'estudiar-ne l'estructura.

Uso de Base de Datos Operacionales	Uso de Data Warehouse
Muchos usuarios concurrentes	Pocos usuarios concurrentes
Consultas predefinidas y actualizables	Consultas complejas, frecuentemente no anticipadas.
Cantidades pequeñas de datos detallados	Cantidades grandes de datos detallados
Requerimientos de respuesta inmediata	Requerimientos de respuesta no críticos

Les aplicacions per a suport de decisions basades en un data Warehousing poden fer més practica i fàcil l'explotació de dades per a una major eficàcia del negoci, que no s'aconsegueix quan s'usen només les dades que provenen d'aplicacions operacionals (que ajuden en l'operació de l'empresa en les seves operacions quotidianes), en els que la informació s'obté realitzant processos independents i a vegades complexos.

Un Data Warehouse es crea a l'extreure dades des d'una o més bases de dades d'aplicacions operacionals.

La dada extreta es **transformada** per a eliminar inconsistències i resumir si es necessari, i després carregar-les en el data Warehouse. El procés de transformar, resumir i combinar els extractes de dades ajuden a crear l'accés a la informació. Aquest nou enfocament ajuda a les persones individuals de tots els nivells de l'empresa a efectuar la seva presa de decisions amb més responsabilitat.

Bàsicament el Data Warehouse es basa en el sistema de suport de decisions que tingui l'empresa (diferenciat del que extreu i usa les dades extretes del dia a dia de l'empresa per al seu funcionament quotidià).
Aquest sistema s'encarrega bàsicament de les funcions del plantejament, previsió i administració de l'empresa, crítiques per a la supervivència d'aquesta. Son les *funcions basades en el coneixement*.

3.2. Característiques d'un DW:

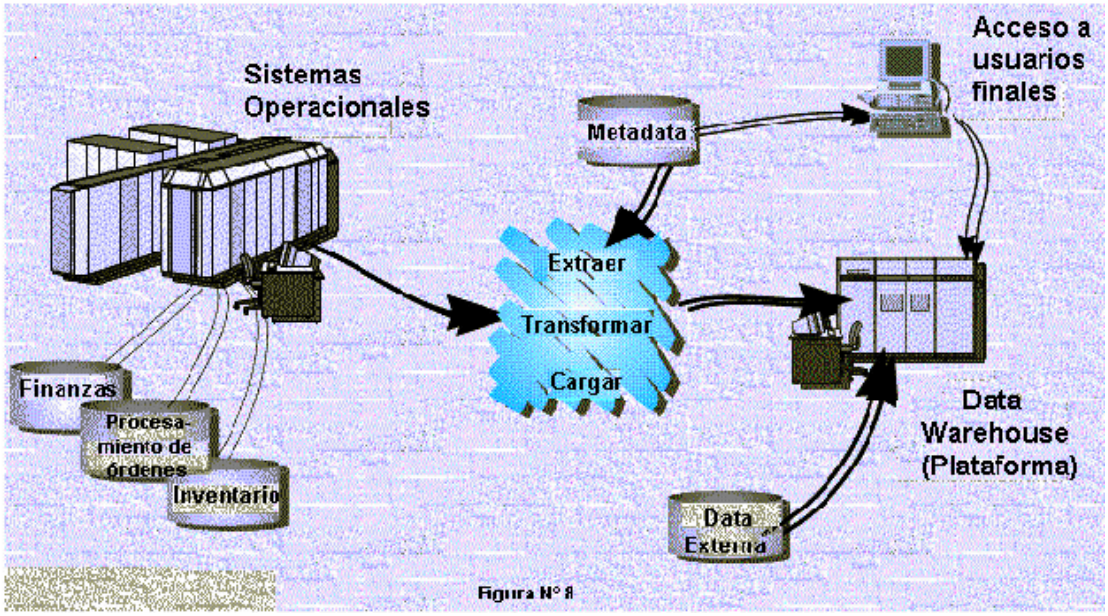
- **Orientat al tema**: el DW s'organitza al voltant de subjectes, de "objectes" que formen una unitat. I cada empresa, cada organització, te uns temes diferents, propis, sobre els quals tenir les dades. Mentre que els dades funcionals poden ser usades o no per l'analista de dades, les del DW només contenen informació que serà usada per al recolzament de decisions.

- **Integrat**: la informació que contenen les DW està sempre integrada, integració supeditada a les decisions dels dissenyadors de les diferents aplicacions. La càrrega de les dades s'haurà de fer tenint en compte la codificació, la mesura dels atributs, les convencions de nomenament i les fonts múltiples. De tot això ens n'hem d'encarregar en el procés de transformació de les dades que integrarà les dades abans d'entrar en el dipòsit.

- **No volàtil**: en contrast amb les dades operacionals que s'actualitzen a cada moment, les dades, un cop "aprovades", transformades, del DW no s'actualitzen, només es carreguen i s'accedeixen per a estudiar-les.

- **De temps variant:** encara que en els processos operacionals –on es volen les dades quasi instantànies- les dades històriques no solen ser de molt us, en el DW la informació por ser demanada en qualsevol moment, ja que es poden usar per a plasmar i avaluar tendències. Per tant les dades (que si son correctes no poden ser actualitzades) solen ser d'un rang de temps llarg.

3.3 CONFIGURACIÓ BÀSICA D'UN DATAWAREHOUSE



- **Sistemes operacionals :** dades extretes dels sistemes d'aplicació operacionals de l'empresa. Son la font principal de dades per al DW.

- **Extracció, transformació i càrrega de dades**. Com abans he esmentat, les dades, un cop aconseguides, han de passar un procés de transformació per a ser integres, amb el mateix format, consistents.

- **Metadata**. La informació sobre les dades son definicions de les dades que agafem. Al crear la metadata també s'ha de tenir en compte la integració per a tenir una coherència entre elles. Normalment consta d'estructures de dades que donen una visió de les dades a l'administrador de dades; de definicions del sistema de registre des d'on es construeix el DW ; d'especificacions de transformacions que passen; del model de dades del DW (elements usats i les seves relacions); un registre de quan els elements s'agreguen al DW, s'eliminen o es resumeixen; els nivells i mètode de sumarització...

- **Accés als usuaris finals.** La interfície amb l'usuari final ha de ser fàcil (d'aquí l'ús de interfícies gràfiques) i tenir les opcions que requereixi cada tipus d'usuari. Existeixen diferents eines per a aquesta tasca.

- **Plataforma del DW.** Normalment un servidor de base de dades relacional on es guarden les dades carregades en el DW. Hi ha que tenir en compte que el dipòsit anirà creixent.

- **Dades externes.** No hi ha que obviar la importància de les dades externes, accessibles per mitjà de computadora en línia o via Internet.

4. OLAP I OLTP. Diferents magatzems

4.1. Resum de les principals diferencies dels magatzems tipus OLTP i OLAP

OLTP: Bases de dades transaccionals, pròpies o incorporades
OLAP: BBDD Data Warehouse d'anàlisi.

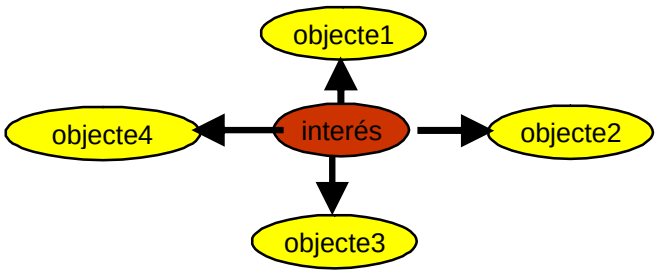
Característica	OLTP	OLAP
Tamany BBDD	GigaBytes	Giga a TeraBytes
Origen Dades	Intern	Intern i Extern
Actualització	On-line	Batch
Períodes	Actual	Històric
Consultes	Predictibles	Ad Hoc
Activitat	Operacional	Analítica

Aquestes diferencies fan que no sigui possible la convivència en una única BBDD dels entorns OLAP i OLTP:

- Pèrdua del rendiment de l'entorn OLTP
- Falta d'integració entre diferents aplicacions OLTP
- Tecnologies de BBDD sense capacitat per a suportar aplicacions OLTP
- Incorporació de dades externes difícilment aplicables a la BBDD OLTP
- Distribució de les dades no adequada per anàlisis OLTP

4.2. BBDD OLAP

L'anàlisi de les dades se sol basar en un model simplificat d'estrella, o més genèricament, de floc de neu (snowflake), que relaciona els fets amb els agents de negoci (dimensions).



on el objecte que volem estudiar es el interès, i ho fem extraient el coneixement que ens donen els diferents objectes amb el qual està relacionat.

La relació entre les taules relacionals i taules de fets i dimensions, es porta a terme mitjançant un Diccionari de Dades, el qual defineix cada element del negoci en base a les taules i camps físics.
Tipus de BBDD :

- BBDD Relacional
- BBDD Multidimensional
- BBDD Híbrida
- BBDD OLAP (BBDD Relacional amb funcionalitat OLAP)

OLTP	***OLAP***
Orientada a transacciones	Orientada a conceptos
Detallada	Sumarizada
Actualizada en línea	Representa valores a un tiempo
Usuarios a nivel operativo	Usuarios a nivel gerencial
Corre en base a repeticiones	Corre heurísticamente
Muy sensitivo al desempeño	Poco sensitivo al desempeño
Accesa unidades a la vez	Accesa conjuntos de unidades a la vez
Orientado a una operación	Orientado a análisis
Estructura estática	Estructura flexible
Sin redundancia	Con mucha redundancia
Alta probabilidad de acceso	Modesta probabilidad de acceso
Administrada como un todo	Administrada por partes
Información bruta (datos)	Información procesada (información)
Actualizada en línea	Actualizada en batch
Muchas tablas con pocas columnas	Pocas tablas con muchas columnas

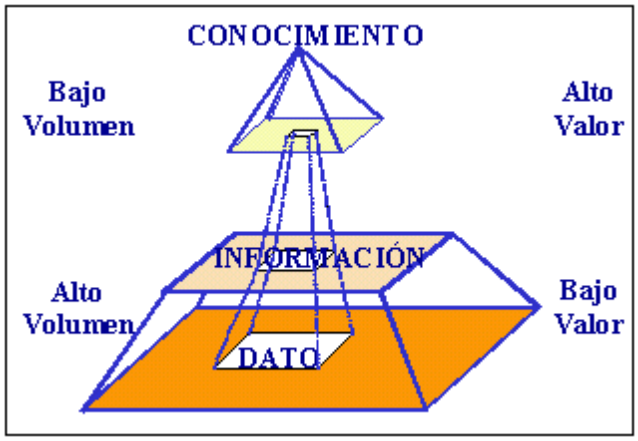
Diferencies MD vs OLAP:

1. OLAP: procés deductiu que permet verificar si certes hipòtesis que fa l'operadors son certes o no.
2. MD: en lloc de verificar patterns de comportament, els descobreix.

5. INTRODUCCIÓ AL KNOWLEDGE DISCOVERY in DATABASES (KDD)

"Proceso no trivial de identificación en los datos de patrones válidos, novedosos, potencialmente útiles, y finalmente comprensibles". Fayyad et al, 1996

A vegades s'usa el KDD com a sinònim de Mineria de Dades. Per a d'altres la Mineria es només un dels passos involucrats en el KDD, que tal com el seu nom deixa clar es la descoberta del coneixement en les bases de dades.



Realment el KDD involucra les etapes de preprocessament de dades, mineria de dades, avaluació dels patrons descoberts i presentació del coneixement.

El conjunt de processos no trivials que possibiliten la identificació de nous patterns en les dades de les bases de dades (vàlids i potencialment utilitzables) són:

1. Adquirir i seleccionar conjunt de dades sobre les que es treballarà.
2. Validació de dades, integració, preprocesament i transformació de dades inicials
3. Elecció de algoritmes de MD
4. Interpretació i visualització de dades
5. Verificació i test de resultats, tuning de models
6. Us i manteniment del "coneixement" generat

I les característiques que han de presentar les Bases de Dades que s'usaran en un KDD:

1. Habilitat per a accedir a variades fonts de dades
2. Accessos online/offline
3. Models de dades : models no estendards
4. Tipus de atributs : a vegades les eines usades per a KDD (MD) presenten restriccions en els tipus de atributs a manejar en la BD.
5. Llenguatge query : en KDD via interfície gràfica (GUI)
6. El tamany de la BD es important a l'hora d'elegir les eines de KDD, per a obtenir bones respostes

6. DSS: Sistemes d'ajuda de Presa de Decisions (Decision Support Systems)

" procés de dades interactiu i un sistema de representació visual (entorn gràfic) que es usat per a ajudar en el procés de presa de decisions"

Les noves opcions ofertes per les Tecnologies de la Informació i Comunicació (TIC) impacten de forma directa sobre la direcció de qualsevol organització, exigint un mode diferent d'actuació si es vol continuar sent eficaç i eficient. Les organitzacions han de substituir el seu sistema tradicional de pensament (centrat en la resolució de problemes a curt plaç i en el pronòstic de futur sobre la base del passat) i deixar pas a un altre sistema de caràcter més estratègic.

L'estratègia de l'organització ha de contemplar, com requisit indispensable per al seu èxit, la creació i manteniment de una base tecnològica adequada que sostingui el procés decisori, permetent-li aconseguir els avantatges competitius desitjat, a més de crear les majors sinergies favorables per a l'organització.

Els sistemes de suport a les decisions (DSS) son la tercera generació de les aplicacions basades en computadores.
L'ús mes freqüent de la primera generació era el processament de transaccions, l'assistència per la presa de decisions quedava en mans de científics e investigadors operatius, que generaven models estructurats, usant les computadores només com a poderosos auxiliars per al càlcul.
Així els sistemes DSS es converteixen en el complement de les capacitats del ser humà, usant la potencia que adopta la informàtica per al processament de dades.

Realment el DSS és un procés de dades interactiu i un sistema de representació visual (entorn gràfic) que es usat per a ajudar en el procés de presa de decisions i ha de reunir les següents característiques:

- Ser lo suficient senzill per a que el pugui utilitzar el que ha de decidir en persona.
- Ha de mostrar la informació en format i terminologia familiar per a l'usuari.
- Ser selectiu en la seva provisió de informació (evitant sobrecarregar a l'usuari)

L'ús de les DSS ha de permetre en una organització:

- millorar el procés de presa de decisions,; proporcionant informació que actualment no existeix, com també millores en l'accés (la visualització, l'anàlisi de la informació). A més es proporcionen millores en el procediments deductius a partir de la informació amb la que disposem, oferint millor explicació a tercers sobre les decisions preses.

6.1. CARACTERÍSTIQUES DELS DSS

- El DSS ha de ser tan "especialitzat" sobre el problema de l'usuari com sigui possible.
- El DSS ha de tenir en compte la problemàtica que envolta els usuaris potencials (la falta de cultura informàtica i la falta de sistematització en els plantejaments dels problemes a resoldre pels directius).
- Els DSS han ser gestionats per experts que entenguin el que els usuaris volen fer.
- Els DSS han de ser controlables pels usuaris, que han de ser capaços de especificar el que volen.
- Els DSS deuran de poder ser capaços d'utilitzar qualsevol dada, model, disciplina eina o tècnica de presentació visual i en definitiva tot allò que faciliti la presa de la decisió.

7. MINERIA DE DADES (MD)

Un Sistema Data Mining nos permite analizar factores de influencia en determinados procesos, predecir o estimar variables o comportamientos futuros, segmentar o agrupar ítems similares, además de obtener secuencias de eventos que provocan comportamientos específicos.

7.1. Introducció a la Mineria de Dades

La Mineria de Dades -l'extracció d'informació oculta y predictible de grans bases de dades-, és una novedosa tecnologia amb molt potencial per a que les companyies puguin extreure la informació més importat de les seves Bases de Dades amb informació (Data Warehouses, XX).
Gràcies als sistemes informàtics i a les eines de la Mineria de Dades es resol el procés feixuc i llarg de respondre a preguntes que són difícils i llargues de fer-ho. Les eines de la MD prediuen futures tendències i comportaments, permetent en els negocis prendre decisions proactives.
L'usuari tracta d'obtenir una relació de les dades que tinguin repercussions en el seu negoci.

Aquestes eines busquen en la base de dades (moltes vegades amb enormes quantitats de dades) patrons ocults, trobant informació predictible.
Es a dir, bàsicament el procés de Mineria de Dades extrau els coneixements útils i utilitzable guardats en les bases de dades de les empreses.
Es una de les eines que tenim actualment al nostre abast per a poder contestar preguntes prèvies a partir d'unes dades donades.

Altres de les eines i les seves diferències són:

Tipo de Herramienta	**Pregunta básica**	**Modelo de Salida**	**Usuario típico**
Consulta y Reporte	¿Qué sucedió?	Reportes de ventas mensuales; histórico de inventario	Necesita data histórica puede tener aptitud técnica limitada
Procesamiento analítico en línea (OLAP)	¿Qué sucedió y por qué?	Ventas mensuales vs. Cambios de precio de los competidores	Necesita ir de una visión estática de los datos a "slicing and dicing" técnicamente astuto
Sistema de Información Ejecutiva (SIE)	¿Qué necesito conocer ahora?	Libros electrónicos; Centros de comandos	Necesita información resumida o de alto nivel puede no ser técnicamente astuto
Data mining	¿Qué es interesante? ¿Qué podría pasar?	Modelos predictivos	Necesita extraer la relación y tendencias de la data ininteligible técnicamente astuto.

Per tant les finalitats d'un DataMining ens permet analitzar factors d'influència en determinats processos, predir o estimar variables o comportaments futurs, segmentar o agrupar ítems similar, i/o obtenir seqüències d'events que provoquen comportaments específics.

7.2. Fonaments de la Mineria de Dades

Els negocis van ser emmagatzemats per primer cop en les computadores, es va millorar l'accés a les bases de dades y les tecnologies van permetre que usuaris naveguessin a temps reals entre les dades… la Mineria de Dades podia ser possible.
La Mineria de Dades aprofita aquestes premisses per a evolucionar, y tres tecnologies fan possible que estigués llest per a la seva aplicació en la comunitat de negocis:

- Recol·lecció masiva de dades
- Potents computadores amb multiprocessadors
- Algoritmes de Mineria de Dades

Les dades de negocis han evolucionat a informació de negocis.
Les àrees d'estadística, intel·ligència artificial y aprenentatge de màquines (entre d'altres) han treballat força per a la Mineria de Dades fent que sigui possible per a entorns Data Warehouse Actuals.

7.3. Finalitats de la Mineria de Dades

Un sistema de Mineria de dades ha de ser capaç de:

1. descobrir les dades en forma resumida, donant les principals propietats estadístiques
2. visualització gràfica de les dades
3. descobrir potencials relacions entre les seves dades
4. construir models predictius, en base als patterns trobats
5. verificar els models construïts

La Mineria de Dades no descobreix solucions automàticament sense guia
Es necessari comprendre les tècniques usades per a poder fer un bon ajust de paràmetres per a optimitzat la precisió dels algoritmes utilitzats.

Freqüentment les dades a ser tractades s'extrauen de DW i s'analitzen des d'un DM o des de data mart. PERO *no es imprescindible* la existència de DW per a que existeixi una MD.

Algunes aplicacions de la Mineria de Dades

- Detectar característiques de clients ("profiling")
- Detectar fraus (targetes de crèdit, telecomunicacions)
- Prediccions: demanda de productes, efectivitat de medicaments, risc de crèdits
- Classificació (reconeixement de patrons de imatges): p.ex. cossos celests.
- Elaboració de estratègies de marketing

- Comerç electrònic: sistemes de recomanació, optimització d'inventaris (text i web mining)

En aquest cas a partir de una base de dades treure'm les característiques dels potencials clients que alhora ens serviran per a elaborar les estratègies de marketing.

7.4. L'abast de la Mineria de Dades

Amb bases de dades de suficient tamany i qualitat el procés de buscar valuosa informació que té el Mineria de Dades dona als negocis oportunitat de disposar de les següents capacitats:

- **Predicció automatitzada de tendències i comportaments**: Data Mining automatitza el procés de trobar informació predictible en grans bases de dades. Preguntes que abans requerien un intens anàlisis manual, ara poden ser contestades directa i ràpidament des de les dades.
- **Descobriment automatitzat de models prèviament desconeguts**. Les eines de Mineria de Dades escombren les bases de dades e identifiquen models prèviament amagats en un sol pas.

7.5. Per què usar Mineria de Dades?

Tornant a fer incapeu, la Mineria de Dades és una tècnica d'intel·ligència artificial que permet facilitar l'anàlisi de les dades.
Les eines del DM recullen dades detallades de transaccions per a desenterrar patrons.
I generalment els resultats generen extensos reports o se les analitza amb eines de visualització de dades descobertes.

- La Mineria de Dades contribueix a la presa de decisions tàctiques i estratègiques proporcionant un sentit automatitzat per a identificar informació clau des de columnes de dades generats per processos tradicionals i de e-Business.

- Permet als usuaris donar prioritat a decisions i accions mostrant factors que tenen un pes en un objectiu, quins segments de clients son despreciables i que unitats de negoci son sobrepassats i perquè.

- Proporciona poders de decisió als usuaris del negoci que millor entenen el problema i l'entorn i es capaç de mesurar les accions i els resultats de la millor forma.

- Genera Models Descriptius: en un context d'objectius definits en els negocis permet a empreses, sense tenir en compte la indústria o el tamany, explorar automàticament, visualitzar i comprendre les dades i identificar patrons, relacions i dependències que impacten en els resultats finals del compte de resultats.

- Genera Models Predictius: permet que relacions no descobertes e identificades a través del procés del Data Mining siguin expressades com regles del negoci o models predictius. Aquests ouputs poden comunicar-se en formats tradicionals per a guiar l'estratègia i planificació de l'empresa.

8. MINERIA DE DADES VS ESTADISTICA

Per als matemàtics i per als analistes de l'empresa (en aquest cas Motorpress Ediciones) la introducció d'una nova manera de treballar amb les dades entra amb escepticisme. Sempre treballant amb l'estadística aquesta secció intenta conciliar i presentar la mineria de dades a aquest sector de treballadors.

Per començar, les dos ciències tenen el mateix objectiu: millorar la presa de decisions mitjançant un coneixement de l'entorn. Aquest entorn el faciliten les dades emmagatzemades en la companyia, quantitatius o qualitatius i mitjançant informació de terceres empreses.

La Mineria de Dades és millor que l'Estadística en els següents supòsits:

- Les tècniques estadístiques es centren generalment en tècniques confirmatòries, mentre que les tècniques de Mineria de Dades són generalment exploratoris.

Per tant quan el problema al que pretenem trobar resposta es refutar o confirmar una hipòtesis, podem usar ambdues ciències. Però quan l'objectiu es només exploratori (per a concretar un problema o definir quines son les variables més interessants en un sistema d'informació) sorgeix la necessitat de delegar part del coneixement analític de les empreses en tècniques d'aprenentatge (intel·ligència artificial), usant la Mineria de Dades. Per tant el Data Mining s'usarà quan no tinguem supòsits de partida i pretenguem buscar el coneixement nou i susceptible de proporcionar informació nova en la presa de decisions.

- A més gran dimensionalitat del problema el Data Mining ofereix millors solucions.

Les tècniques de Data Mining son menys restrictives que les estadístiques. Una vegada trobat un punt de partida interessant i disposats a usar l'anàlisi estadístic en particular, pot succeir que els elements de dades no donin els requeriments de l'anàlisi estadístic. Llavors les variables seran examinades per a determinat que el tractament permet adequar-les a l'anàlisi, no sent possible o convenient en tots els casos. Aquí també destaca el Data Mining, posat que es menys restrictiu que la estadística i permet ser usat amb els mínims supòsits possibles.

- Quan les dades de l'empresa son molt dinàmiques les tècniques de la Data Mining incideixen sobre la inversió i la actualització del coneixement del negoci.

Un magatzem de dades poc dinàmic permet que una inversió en un anàlisis estadístic quedi justificat (personal, metodologia rígida i resposta a preguntes molt concretes), donat que les conclusions van a tenir un cicle de vida llarg. Però en un magatzem molt dinàmic, les tècniques del Data Mining permeten explorar canvis i determinar quan una regla de negocia ha canviat. Permet abordar diferents qüestions a curt/mig plaç.

L'anàlisi estadístic es mes adequat que el Data Mining quan:

- L'objectiu de la investigació es trobar la casualitat.

Si es pretén determinar quines son les causes de certs efectes, s'han d'usar tècniques de estadística. Les relacions complexes de les tècniques de data Mining impedeixen una interpretació certesa de diagrames causa-efecte.

- Es pretén generalitzar sobre poblacions desconegudes en la seva globalitat.

Si les conclusions han de ser extensibles a altres elements de poblacions similars han d'usar-se tècniques d'inferència estadística. En Mineria de Dades, es generaran models i després hauran de validar-se en altres casos coneguts de la població, utilitzant com significació l'ajust de la predicció sobre una població coneguda.

Però s'ha de destacar també que **ambdues perspectives es complementen** en la tasca d'obtenir coneixement inèdit en els nostres magatzems de dades o donar respostes a qüestions concretes de negoci, i no són excloents l'un de l'altre.

La metodologia d'un projecte de Data Mining ha de contenir referències a l'estadística en dos parts destacables del procés:

- Preparació de les dades i aproximació a les variables d'estudi.
- Desenvolupament del projecte i possible generació de hipòtesis a refutar amb una metodologia i tècnica estadística.

9. PROCESSOS EN LA CERCA DEL CONEIXEMENT

Dades:

- tradicionalment una taula en ASCII
- tendència a Magatzems de Dades que estan optimitzades per a procés analític
- eines de KDD poden incloure mecanismes per a emmagatzemar dades i accedir a dades

Selecció:

- selecció de conjunt o subconjunt de bases de dades
- selecció de subconjunt de variables a usar en MD
- selecció de mostres de dades (instàncies)

Pre-processament:

- neteja de dades i preprocesament
- eliminació de soroll i casos extrems si es necessari
- arreglar les dades que falten i les desconegudes

Transformació:

- transformació al format necessari per l'algoritme específic de Mineria de Dades

Mineria de Dades:

- Búsqueda de patrons d'interès en una forma partícula de representació, que poden expressar-se com un model o com un patró que expressa certa dependència de dades
- El model:
  - la seva funció (classificació, regressió, clustering,...)
  - forma de representar-lo (funció lineal, conjunt de regles,...)
- Criteri de preferència (Quin model es millor? Quin conjunt de paràmetres del model?)
- Estratègia de búsqueda

Interpretació/Avaluació:

- Interpretació dels patrons descoberts, pot beneficiar-se moltíssim usant visualització
- Es poden esborrar patrons redundants o irrellevants
- Els patrons poden comparar-se amb el coneixement prèviament emmagatzemat (o extret)

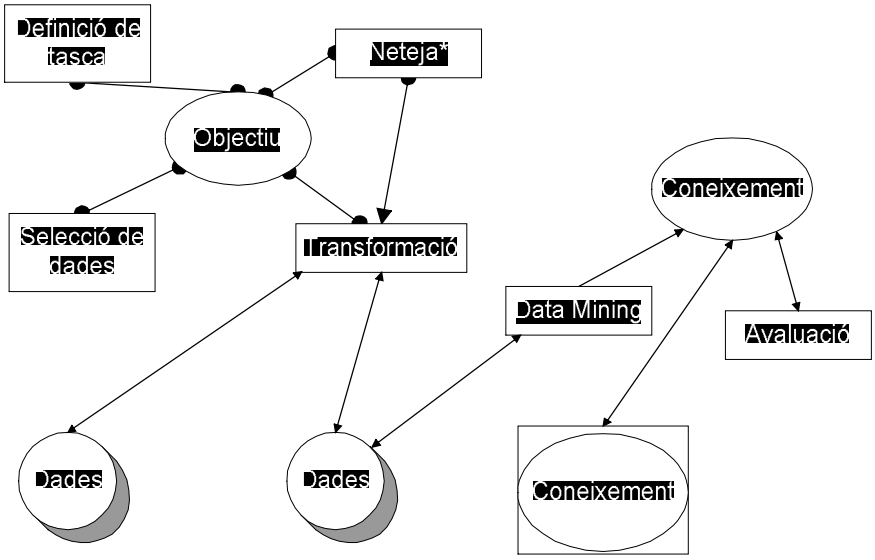
Coneixement:

- Realitzar accions
- Incorporar al coneixement descobert en un sistema
- Documentar el coneixement i reportar-lo a les persones interessades

D'aquesta manera extrèiem el coneixement a partir de les dades.

(Nota: amb aquesta descripció dels processos ens reafirmem en que la Mineria de Dades es només una petita part de la gran tasca, la cerca del coneixement (KD)).

10. CICLE DE VIDA DEL PROCÉS DE DESCOBRIMENT



1.A. Definició de tasca.

Per a definir quina es la tasca principal del projecte hem de definir quin és l'objectiu que volem assolir del treball que anem a dur a terme, quines son les preguntes que pretenem respondre. Aquest objectiu el relacionarem amb una tasca, un model per a treballar les dades (no ha de perquè reduir-se tot l'objectiu a un sol model). Per exemple*:

* Com que les tècniques que veurem després, que s'usen per a poder aconseguir l'objectiu que busquem, es poden combinar i ser usades de diferents maneres, encara no hi ha acord de com es poden dividir segons l'objectiu buscat. Jo agafo com a base el vist en l'assignatura Mineria de Dades d'ETIG.

- Predir. Si volem predir algun comportament podríem fer servil els models de regressió o els arbres de decisió.
- Classificar. Quan volem saber les característiques que diferencien els components d'un grup dels d'un altre serà millor usar arbres de decisió (com els arbres de Classificació i Regressió CART), les xarxes neuronals artificials o les simples regles de classificació (antecedente Þ consecuente on l'antecedente es una llista d'una o varies variables amb rangs associats).
- Semblança. Buscar similituds entre les dades que tenim podem fer-ho mitjançant models associatius o els models d'agregació (clustering)
- Descriure. Quan el que busquem es saber com d'associades estan les variables, quines tenen més pes en que també estiguin presents altres o si al contrari, la seva aparent relació es casual, lo millor serà usar regles d'associació o models descriptius con les xarxes bayesianes.
- Explicar. Si el que volem es saber el perquè d'un comportament. Per a això els models explicatius com les xarxes bayesianes seran ideals.

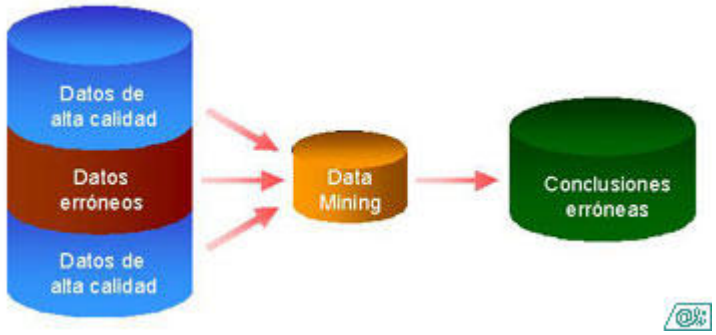
Dues agrupacions d'objectius més que he trobat en un treball de Data Mining són:

◊ associacions; dependències; classificació; segmentació; tendències i regles generals.

◊ predicció de series temporals; classificació de patrons; agrupació de característiques; lògica difusa, extracció de regles i coneixement; optimització de processos; extracció des de Internet (o Web Mining) en quant al comportament dels navegants; modelització de sistemes complexos.

1.B. Selecció de dades.

No totes les dades que ens pot proporcionar una empresa ens són útils i necessaris per a aconseguir el nostre objectiu. Al contrari, si no seleccionem correctament les nostres dades per a l'anàlisi podem no treure conclusions correctes.
La tecnologia del DataWarehouse que emmagatzema les dades de les diferents àrees de l'empresa, vista anteriorment, serà ideal com a proveïdor d'aquestes dades. Algunes de les seves característiques (dades del negoci per a la informació, orientat al subjecte, dades actuals+històriques, detallades i també resumides, i estable) ens donen fiabilitat a l'hora de seleccionar les dades del DW.
Aquesta es la primera fase d'aconseguir unes dades de qualitat, que conclourà amb les tasques de neteja i transformació.



Dades objectiu:

Ara ja tenim unes dades sotmeses a uns objectius que pretenem aconseguir.

2. OBJECTIU

Un cop definit l'objectiu de l'empresa i tenir unes dades elegides amb les quals treballarem, tenim que passar a fer que aquestes dades estiguin en format i manera òptima per al seu ús. Per tant un cop tenim l'objectiu clar passarem a netejar i transformar les dades per a poder tenir-les a punt per al procés de Data Mining.

3.A. Neteja (si es necessari)

Bàsicament (com veurem en l'apartat dedicat a la preparació de les dades, la neteja tractarà:

- l'eliminació general de les duplicitats en les dades
- sincronització de les referències a objectes o categories
- correcció d'errors o dades no vàlides:
    - dades incompletes

  - dades incorrectes o inconsistents
- prescindir de dades no òptimes per al nostre objectiu:
  - dades esbiaixades
  - dades envellides

Dades preprocessades:

Ara ja tenim les dades que creiem majoritàriament útils, encara que aquesta neteja potser ha estat mínima si el procés d'emmagatzematge d'allà on les hem tret (suposant que només sigui una font), ja havia realitzat aquesta tasca.
Però encara necessitarem treballar-les per a poder aplicar els algoritmes corresponents.

3.B. Transformació

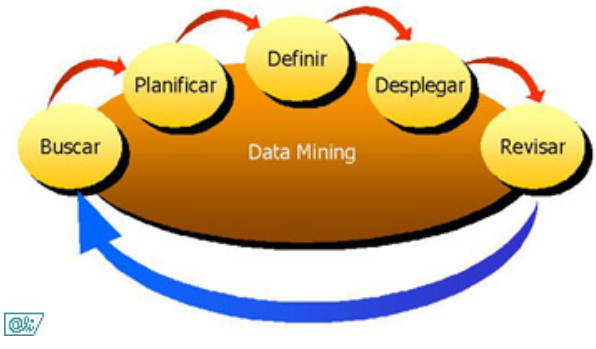
Aquí el que es farà es adequar l'estructura de les dades. Entre altres transformacions podem efectuar:

- Categorització de les dades
- Numerització dels valors categòrics
- Jerarquització d'atributs
- Eliminació de files i/o columnes (reducció de la dimensionalitat)
- Simplificació de valors

Dades transformades:

Ara ja tenim les dades processades per a poder començar les tasques pròpies de la Mineria de Dades. Hi ha que tenir en compte que els passos que acabem d'efectuar poden usar la major part del temps d'un KDD (fins i tot un 70% o un 80% d'aquest temps).
Les dades tenen ja la qualitat adequada per a començar el procés de MD.

4. Data Mining



Ja estem a en disposició de construir els models necessaris per a extreure coneixent.
Per a fer-ho tornarem a revisar l'objecti, i ara si, a partir d'un model o de cap, es buscarà un complet que tingui prou qualitat per a arribar a l'objectiu prefixat (es a dir, contestar les preguntes que ens havíem fet a l'inici de tot).
El procés de Mineria de Dades, el per què i les finalitats del qual ja hem vist i alguns models de la qual també veurem detalladament podria dividir-se en tres grans fases:

. el modelatge, que té per definició allò que esmentava abans, el construir un model per a produir una resposta a una pregunta tenint unes respostes per a la mateixa pregunta en altres situacions. El modelatge es sol realitzar fent reiterats feed-backs.
. la constatació de que el model té prou qualitat (i si no tornar a la fase anterior)
. la presentació del model a l'usuari mitjançant tècniques de visualització que pugui entendre

Patrons

Ara de les dades ja hem extret patrons reconeguts.

5. CONEIXEMENT

6. Avaluació

Podem desfer-nos del redundants i també comparar-los amb el coneixement previ que tinguem emmagatzemat, i mirar el que volen dir aquests patrons en relació amb les preguntes que volíem contestar.

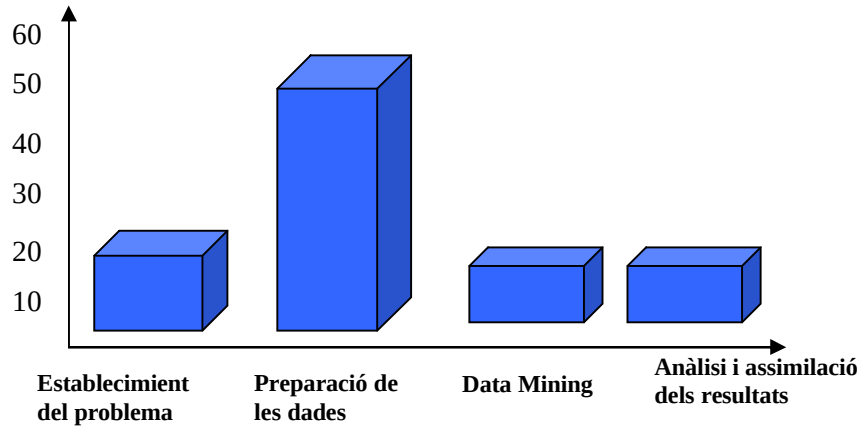
7. CONEIXEMENT

A partir d'aquí es pot fer una altra vegada el treball de cerca de coneixement, ja que aquest pot ser (i hauria de ser) un procés constant.

11. PREPARACIÓ DE LES DADES I TASQUES PRÈVIES A L'ANÀLISI.

"si no ho fem correctament és molt probable que posteriorment puguem treure patrons no representatius"

Com hem vist abans, el preparar les dades per a que puguem aplicar els mètodes i construir els models que ens donin respostes, serà el procés que s'emportarà la major part del temps quasi sempre.



Estimació aproximada de la distribució del temps

El que hauríem de fer per a preparar correctament les dades per a un procés de Mineria de Dades seria:

- Analitzar les dades per a descobrir inexactituds, anomalies i altres problemes.
- Transformar les dades per a que siguin precises i coherents, eliminant els problemes anteriors.
- Assegurar la integritat referencial
- Validar les dades per a realitzar les consultes de prova
- Produir la metadata (una descripció del tipus de dades, format i el significat relacionat al negoci de cada camp.
- Finalment, el crear una documentació del procés complet ajudarà a que es puguin ampliar, modificat i arreglar les dades en un futur amb més facilitat.

Encara que els processos de documentació i metadata no es consideren com a tasques prèvies a l'anàlisi en molts dels articles consultats, l'experiència demostra que resulten molt útils quan volem fer una Gestió del Coneixement continuada, més enllà d'un treball de Mineria de Dades puntual.
La resta de les tasques les podríem fer entrar en les fases de neteja i transformació del procés de Búsqueda de Coneixement vist anteriorment. Ja que trobar unes eines (encara que si que existeixen) que donin un bon resultat al fer aquestes tasques, anem a veure-les amb una mica més de deteniment, fent abans una breu introducció al tipus de dades que usem normalment.

11.1 Tipus de dades

- Quantitatives:
  - Discretes
  - Continues
- Qualitatives (o Categòriques):
  - Nominals (nombren a l'objecte al qual es refereixen)
  - Ordinals (es pot establir un ordre en els seus valors)

* Nota: hi han varies maneres d'exposar els tipus de variables. Aquesta és una de les més senzilles.

11.2. Neteja de dades

El format de les dades, els atributs, ha de presentar una correcta qualitat. Per a fer-ho s'ha de tenir en compte:

a) **les duplicitats en les dades.** Es poden repetir columnes o files que són el mateix. Per exemple:

DNI	Nom	Cognoms
47628032	Cristina	Colell Ventepani
47628032	Cristina	Collell Ventepani

Nom i Cognoms	Edat	Anys
Cristina Collell Ventepani	23	23

Identificador	Nom	Cognoms
00012	Cristina	Collell Ventepani
00012	Daniel	Roselló Roselló

Els dos primers casos poden passar per exemple en una fusió de bases de dades o (en el primer cas) si es dona dos cops la informació i una de elles es fa malament (voluntària o involuntàriament).
En el tercer cas una duplicació de identificador pot ser un error o pot ocórrer també en una fusió de bases de dades si aquest no es clau en el resultat.
Per a arreglar-ho s'ha de fer el difícil treball d'identificar quin ha estat l'error i quines són les dades correctes.

b) correcció d'errors o dades no vàlides:

1) **dades incompletes**. Quan un camp no era obligatori per a donar d'alta una tupla ara ens farà falta si no s'ha completat. Per tant ara tocarà o emplenar-la amb un "possible" valor o, si es possible, amb una mitja de la resta de valors de l'atribut introduïts. També es pot introduir un valor que no vulgui dir res i que no es prengui en consideració.
Hi ha que esmentar que hi han programes que mitjançant algoritmes posen un valor als camps incomplets basant-se en probabilitats.
Alguns exemples d'aquest error podrien ser:

Carrer	Numero	Pis	Porta

Mesa	12	2	
Roig	2	3	A
Adrià	31		

Aquí podríem posar un 0 als valors que falten, o un caràcter qualsevol si en els camps de porta no es pot posar un caràcter numèric.

Cognoms	Edat	Fills
Sanchez Glor	14	
Asiet Roig	32	
Martí Martí	28	2

Es podria usar una mitja o un valor que per a nosaltres no signifiqui res (i posar-ho en la memòria per a posteriors anàlisis)

2) **dades incorrectes o inconsistents**. Poder també relacionat amb el cas a), les dades incorrectes poden ser resultat d'una introducció lliure o un error en l'introducció dels valors.

El cas propi es quan el valor introduït no es un valor possible (normalment) com el tenir fills als 2 anys (en aquest cas la tasca serà trobar si la dada incorrecta es l'edat de la persona o la dels fills.

Nom	Edat	Fills
Eva	5	2

Num matricula 04/05	Nom	Cognoms
010	Cristina	Collell
010	David	Castellà

[..]	Carrer	Codi Postal
[..]	Amposta	43870
[..]	La Ràpita	43870

Com en el cas a) en aquest cas serà necessari trobar l'error i saber quin valor es correcte.

c) prescindir de dades no òptimes per al nostre objectiu:

1) **dades esbiaixades.** Per a mirar si les dades de les que disposem son esbiaixades hem de tenir en compte quin tipus de pregunta estem tractar de contestar o quina tasca pretenem dur a terme.

Per exemple, si volem mirar quantes hores de programes de "premsa rosa" es veuen a la setmana pot ser un error agafar només gent jove, o gent que no treballa fora de casa. Es a dir, quan agafem sense voler un grup de gent determinada que no es correspon al rang que volem estudiar per les seves característiques determinades.

2) **dades envellides.** Quan hi ha camps d'edats, adreces, nivell d'estudis o dades que varien de valor durant el temps pot ser que si no s'han fet bones actualitzacions en les bases de dades aquests errors es reflecteixin en les dades que anem a tractar.

Nom	Edat	Categoria	Fills
Ana	22	Jove	0

Jordi	35	Adult	3

En aquesta taula podríem tenir envellits els atributs d'edat, i independentment (ja que podem tenir l'edat actualitzada si la relacionem amb un camp de data de naixement) també la categoria. El nombre de fills podria també estar envellit.
Hem de tenir en compte el saber trobar aquestes dades envellides i procedir a eliminar l'error (actualitzant o eliminant).

d) sincronització de les referències a objectes o categories:

Encara que s'hagi dut a terme una bona tasca d'emmagatzemament en els Data Warehouse, pot ser que el fet de treure dades de diferents fonts o algun error previ, dugui a que també haguem de verificar aquesta part.

Cognoms	Ciutat	Professió
Solà Roig	Sant Carles de la Ràpita	Carter
Andreu Martí	La Ràpita	Funcionari

Aquí la mateixa ciutat és anomenada de dos maneres diferents, a més el funcionari podria ser carter, però haver elegit un càrrec més general al no trobar l'opció.
També pot passar en els títols dels camps, però aquest detall hauríem d'haver solucionat en el problema de la duplicació.

11.3. Transformació de dades

Amb les dades ja netes, el procés de transformació sol ser quasi sempre necessari per adequar al màxim el format de les dades al procés al que les volem sotmetre i les dades que esperem obtenir.
D'aquesta manera, a més d'aconseguir unes dades integrades, consistents i consolidades, també hem d'aconseguir que compleixin amb els requisits d'entrada dels algoritmes.

Les transformacions que hem de mirar si són susceptibles de fer-se són:

a) **Categorització de les dades**. Pot ser que ens interessi tenir un atribut categòric en lloc de numèric. Per exemple:

Sou mensual
4.000

Sou mensual
Alt

Mentre que en la primera taula mesurem l'atribut sou mensual amb un tipus numèric continu, en la segona l'hem categoritzat (per ex. en el rang: molt baix, baix, mitjà, alt, molt alt, amb els respectius valors)

b) Numerització dels valors categòrics. Al contrari que abans, farem numèrics les dades categòrics. Igualment però, la majoria de vegades es quedaran en forma de rang, ja que per exemple, el valor alt de l'atribut sou, pot ser que siguin de 3.000 a 5.000€.

c) Eliminació de files i/o columnes (reducció de la dimensionalitat): Aquesta transformació posa de relleu la importància de la búsqueda d'una informació vàlida amb les menys dades possibles.

a) reducció dels registres: la meta es aconseguir un conjunt igualment representatiu que al analitzar pugui donar els mateixos bons resultats que el conjunt original.
b) reducció dels atributs: els atributs que són combinació de altres atributs (o aquests mateixos), o els irrellevants poden eliminar-se, ja que no repercutiran en el resultat final.

d) Simplificació de valors. Hi han alguns valors que podem simplificar, tant per a estalviar temps de processament (com més bits més temps) com per a estalviar espai.
P.ex. els valors d'una hipoteca poden expressar-se en mils d'euros en lloc d'euros.

e) Codificació: ja tractat abans, el posar els atributs en la forma apropiada per a poder ser tractats pels algoritmes és molt important. Pex. deduir la província si tenim la ciutat però necessitem aquesta primera o posar l'altura en metres si la tenim en centímetres.

f) Normalització de dades: Normalitzar les dades vol dir posar dins d'un rang de valors els equivalents als reflexats en els camps. Resol el problema de les dades esbiaixades, com els valor que surten amb diferencia de la resta (outliners).

- Normalització per intensitat global
  - escala lineal
  - escala logarítmica
- Normalització per ajust de valors
- Normalització per regressió lineal d'intensitats, escala logarítmica

g) Discretització: al discretitzar un atribut el que fem es passar de un conjunt de dades (un atribut continu) a uns valors d'un conjunt (intervals discrets), mitjançant una regla, el que aconseguim es estalviar espai i temps al treballar les dades, a més de ajudar en classificacions inicials.

- Els mètodes poden classificar-se en globals (discretització de tots els valors continus) i locals (discretització de regions de l'espai) i també en:
- No supervisats (independents de les classes):
  - Mètodes de partició en intervals de la mateixa longitud
  - Obtenció d'intervals de discretització d'igual freqüència
  - Per estimació
  - Mètode k-means.
- i mètodes supervisats (prenen en compte les classes)
  - Mètode khi merge
  - Mètodes basats en mesures d'entropia

11.4. Tasques prèvies a l'anàlisi

Encara que ja seria possible passar a aplicar algoritmes sobre les dades eficaçment, encara podem treballar aquestes dades per a treure informació prèvia, de la que potser ens podrem aprofitar després al aplicar els mètodes corresponents.

Alguns blocs d'accions són:

- Mètodes estadístiques i de visualització.
- Treballs amb les dades categòriques:
    - Histogrames
    - "Pie charts" (gràfics de pastis)
    - Distribució de variables
- Treballs amb les dades qualitatives:
    - Mitjana $\text{media aritmética} = \frac{x_1 + x_2 + \cdots + x_n}{n}$
    - mediana: ordenats tots els valors en ordre creixent la mediana es el valor que ocupa la posició central.
    - Variància: $V = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n} = \frac{\Sigma (x_i - \bar{x})^2}{n}$
    - Moda: el valor que apareix amb més freqüència en el conjunt de dades. Si son dos els valors que es repeteixen amb més freqüència el conjunt té dos modes.
    - Plots: caixes on es distribueixen visualment els valors a partir de preses estadístiques.

12. MÈTODES EMPRATS EN MINERIA DE DADES

Vistes ja les funcionalitats que podem requerir en un treball de MD, ara passarem a conèixer en més profunditat les tècniques que podem usar per a aconseguir treure un(s) model(s) que ens serveixin com a resposta.

- No supervisades (no hi ha classes predefinides):
  - Agregació
    - Clustering
      - basat en particions: k-means, k-medoids.
      - Jeràrquics: aglomeratiu, divisori
      - basat en probabilitats
      - ...
  - Associació
    - A priori
    - Regressió multivariant
- Supervisades (hi ha classes predefinides):
  - Predicció
    - Regressió
      - Estadística
      - Establiment de probabilitats
    - Arbres de Predicció
    - Estimador de nuclis
  - Classificació
    - Taules de decisió
    - Arbres de decisió (ID3, C4.5, CART, CHAID, LMDT...)
    - Inducció de les Regles de Classificació
    - Bayesiana
    - Amb exemplars com a base
    - Xarxes neuronals
    - Lògica Borrosa
    - Tècniques genètiques

Nota: Aquí queda l'estructura del com tractaré els capítols que segueixen explicant aquests mètodes. Però els següents quadres donen una estructura i visió diferent de tots els mètodes emprats en Mineria de Dades:

Cuadro 1. Técnicas y algoritmos estadísticos de minería de datos.

Tarea	Técnica	Algoritmos	Fuente
Clasificación	Análisis discriminante	- Discriminante lineal - Función de clasificación - Regla de verosimilitud - Regla discriminante cuadrática - Vecino más cercano† - Regla de Bayes - Regresión logística	Dallas, 2000; Weiss y Kulikowski, 1991.
Agrupamiento	Análisis por agrupamiento	- Vecino más cercano† - Vecino más lejano - Método del centroide - Método del promedio - Varianza mínima de Ward - Selección de simientes	Dallas, 2000; Weiss y Kulikowski, 1991.
Dependencia	Análisis de varianza	– Coeficiente de correlación lineal	Dallas, 2000; Infante, 1990
	Análisis de regresión	– Regresión lineal simple† – Regresión lineal múltiple	Infante, 1990; Dallas, 2000; Draper y Smith, 1966
Series de tiempo	Suavización de curvas	– Promedios móviles simples – Suavización exponencial simple – Suavización exponencial simple de respuesta adaptativa – Promedio móvil lineal – Suavización exponencial de un parámetro (método de Brown) – Suavización exponencial de dos parámetros (método de Holt) – Suavización exponencial cuadrática – Método de Winters	Box y Jenkins 1976; Burés, 1989.
	Ajuste de curvas	– Emparejamiento de expresiones	Guzmán, 1999.

† Algoritmos utilizados en software de minería de datos.

Cuadro 2. Técnicas y algoritmos de IA para la tarea de clasificación.

Técnica	Algoritmos	Fuente
Redes neuronales	- ARTMAP - Backpropagation - Red de función de base radial - Red neuronal probabilístico - Cuantificación del vector aprendizaje	Bigus, 1996.
Arboles de decisión	- ID3	Quinlan, 1986.
	- C4.5	Quinlan, 1993.
	- CART	Berson *et al.*, 2000.
	- CHAID	Berson *et al.*, 2000.
	- CN2	Clark y Boswell, 1991.
Inducción de reglas	- AQ15	Michalski, 1998.
Programación lógica inductiva	- CIGOL	Muggleton y Buntine, 1998.
	- MARVIN	Sammut y Banerji, 1986.
	- PROGOL	Muggleton, 1995.
	- GOLEM	Muggleton y Feng, 1990.
	- MIS	Shapiro, 1983.
	- MFOIL	Dzeroski y Bratko, 1992.
	- FOCL	Pazzani y Kibler, 1992.
	- FOIL	Quinlan, 1996.
	- LINUS	Lavrac *et al.*, 1991.
	- MOBAL	Morik *et al.*, 1993.
	- CLAUDIEN	Dehaspe *et al.*, 1994.

Cuadro 3. Técnicas y algoritmos de IA para la tarea de agrupamiento.

Técnica	Algoritmos	Fuente
Redes neuronales	- Mapeo de caracteristicas Kohonen - Teoria resonancia adaptativa	Bigus, 1996.
Inducción de reglas	- CLUSTER/S	Stepp y Michalski, 1986.
	- CLIQUE	Agrawal, 1998.
	- PARAMETRIZED	Ramkumar y Swami, 1998.
Modelos gráficos probabilisticos	- AutoClass	Cheeseman y Stutz, 1996.
Hipergrafico	- HMETIS - Min-Apriori	Han E. *et.al.*, 1998

Cuadro 4. Técnicas y algoritmos de IA para la tarea de dependencia.

Técnica	Algoritmos	Fuente
Modelos gráficos probabilisticos	– Red bayesiana	Buntine, 1996.
Programación lógica inductiva	– CLAUDIEN	Dehaspe *et al.*, 1994
Inducción de ecuación	– BACON	Rich y Knight, 1994.
Inducción de reglas	– Basic – Cumulate – EstMerge	Srikant y Agrawal, 1995.
	– MultipleJoins – Reorder – Direct	Srikant *et al.*, 1997.
	– Agrawal93	Agrawal *et al.*, 1993.
	– Mannila94	Mannila *et al.*, 1994.
	– AprioriTid – AprioriHybrid	Agrawal y Srikant, 1994.
	– Bayardo99	Bayardo *et al.*, 1999.
	– AprioriUDF	Sarawagi *et al.*, 1998.

Cuadro 5. Técnicas y algoritmos de IA para la tarea de series de tiempo.

Subtarea	Tecnica	Algoritmos	Fuente
Pronóstico	Redes neuronales	– Red de función de base radial – Propagación hacia atrás recurrente – Aprendizaje de diferencia temporal	Bigus, 1996.
Busqueda de patrones	Redes neuronales	– Propagacion hacia atras	Martinez, 1991.
	Induccion de reglas	– Alabeo de tiempo dinamico	Berndt y Clifford, 1996
		– Evento estructura	Bettini C. *et al.*, 1998.
Descubrimiento de patrones secuenciales	Induccion de reglas	– AprioriSome – AprioriAll	Agrawal y Srikant, 1995.
		– GSP	Srikant y Agrawal, 1996.

12.1. AGREGACIÓ

CLUSTERING

Realment sinònim d'agregació, permet fer particions (subgrups o clústers) del grup original. La unió d'aquests subgrups formarà el grup inicial. Un objecte només pot estar en un subgrup.

Els **criteris de similitud** són els que decideixen com es formen els subgrups, per tant seran important la seva bona elecció per a aconseguir un bon resultat. S'expressen en funció de distància d(i,j). Aquesta funció serà diferent depenent del tipus de variables que tractem. Si les dades tenen significats diferents es poden aplicar pesos sobre les variables.

Bases del clustering:

- Els registres similars es divideixen en clusters.
- Donat un conjunt de dades es dona com a resultat una divisió de manera que es divideix la població minimitzant la distancia dels elements de cada subgrup (alta similitud dins de cada classe) i es maximitza la distancia entre subgrups (baixa similitud entre objectes de diferents classes).
- S'ha d'especificar el nombre màxim de clusters, el nombre mínim d'elements en cada cluster, el número d'interaccions i no es necessita l'atribut de sortida al generar clusters que no es coneixen prèviament.
- S'aplica:
    - per a conèixer més les dades (per això hi ha qui ho fa també dins les tasques de preparació de dades).
    - com a preprocés d'altres algoritmes.

SIMILITUDS:

Variables d'intervals

1. Estandarditzar les dades

    1.1. Calcular la desviació respecte a la mitjana

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + ... + |x_{nf} - m_f|)$$

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + ... + x_{nf})$$

    1.2. Calcular la mitjana estàndard:

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

Les distancies s'usen per a mesurar la similitud entre dos objectes:

Distancia de Minkiwski:

$$d(i,j) = \sqrt[q]{(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + ... + |x_{i_p} - x_{j_p}|^q)}$$

on i I j són dos objectes amb p atributs I q es un enter positiu[i=(xi1, xi2,…,xip) i j=(xj1,xj2,…,xjp)].

Si q fos igual a 1, llavors d es la distancia de Manhattan:

$$d(i,j)=|x_{i_1}-x_{j_1}|+|x_{i_2}-x_{j_2}|+...+|x_{i_p}-x_{j_p}|$$

I si q es igual a 2 d es la distancia d'Euclides:

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+...+|x_{i_p}-x_{j_p}|^2)}$$

Propietats de les distàncies:
1. d(i,j) ≥ 0
2. d(i,j)=0
3.d(i,j)=d(j,i)
4.d(i,j) ≤ d(i,h)+d(h,j)

Els mètodes de similitud per a les variables d'intervals també s'usaran (un cop tractades) per a les:

Variables ordenades

Portar el rang de la variable a [0,1] reemplaçant l'objecte i de la variable f:

$$z_{if}=\frac{r_{if}-1}{M_f-1}$$

Variables binàries (poden prendre els valors 0, 1)

Si la variable es simètrica (ambdós valors tenen el mateix pes) usem el coeficient invariant:

$$d(i,j)=\frac{b+c}{a+b+c+d}$$

on a es el nombre de valors que son 1 en les dos, b el nombre de valors que son 1 en i i 0 en j, c el nombre de valors que son 0 en i i 1 en j i d és el nombre de valors que son 0 en les 2.

Si la variable es asimètrica usem el coeficient de Jaccard

$$d(i,j)=\frac{b+c}{a+b+c}$$

Variables escalars no lineals (p.ex exponencials)

- o tractar-les com una numèrica normal
- o realitzar una transformació per a convertir-les en lineals
- o considerar-les com variables ordinals

Variables nominals (les variables poden diferents valors)

- Aplicació simple :

$$d(i,j)=\frac{p-m}{p}$$

on m és el nombre de valors iguals i p el nombre total de variables

- es poden incloure pesos per a donar més rellevància a m.
- es pot crear una variable binària per a cada un dels estats nominals (per exemple, es jove o no).

Mètode numèric: a causa de que treballen sobre dades numèriques per a trobar valors sobre els que decidir, no es poden aplicar, per exemple, en dades categòriques –encara que a voltes existeixen variants-.

⇒ **K-Means**

Donat K:

- es divideixen els objectes en K subconjunts no buits.
- es calcula el centroide: punt mig centre d'un grup potencial
- assignar cada objecte al clúster mirant el centroide tingui més a prop
- i es torna a calcular el centroide i a assignar els objectes fins que no es produeixin variacions.

algoritme:
selecciona k objectes aleatòriament
mentre hi ha canvis **fer**
    assignar cada objecte al cluster més paregut amb el punt mig
    actualitzar el valor de les mitjanes del cluster
fi mentre

En les dades categòriques es pot usar utilitzant en lloc de mitjanes les modes o mesures de similitud. Aquesta variació s'anomena **K-Medoids** i el que fa en lloc d'agafar la mitjana es agafar les modes, els elements representatius,

Mètodes jeràrquics:

No necessiten un nombre de cluster com entrada, però si una condició de finalització.
S'usa la matriu de distàncies com a criteri.
Hi ha dos maneres d'arribar: aglomeratiu (a partir de les dades individuals és van formant grups per similituds, o millor dit, mirant els que són menys diferents) i divisius (a partir del conjunt complet es van format grups més petits).

Mètodes basats en probabilitats:

El que volem es trobar el grup de cluster més probables donades les dades on els objectes tinguin certa probabilitat de pertànyer a un clúster.
Aquest mètode es basa en un model estadístics, la mescla de distribucions (finite mixtures). Una mescla es un conjunt de k distribucions representant k clusters.
Cada distribució ens dona la probabilitat de que un objecte tingui un conjunt particular de parells atribut-valor si formés part realment d'aquest clúster.
Els passos més senzills son:

- donat un conjunt de dades determinar les k distribucions normals:

les mitjanes:

$$\mu = \frac{x_1 + x_2 + \dots x_n}{n}$$

i les variàncies:

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots (x_n - \mu)^2}{n - 1}$$

- calcular les probabilitats particulars de cada distribució:

la probabilitat de que un objecte x pertanyi a un cluster A es:

$$P(A|x) = \frac{P(x|A)P(A)}{P(x)} = \frac{f(x; \mu_A, \sigma_A)P_A}{P(x)}$$

on la funció f es una distribució normal:

$$f(x; \mu_A, \sigma_A) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- eliminem P(x) i normalitzem

Algoritme EM (Expectation Maximization): No sabem de que distribució ve cada dada i no coneixem els paràmetres de les distribucions:

Expectation: el càlcul de probabilitats de les classes o els valors esperats de les classes

Maximization: calcular els valors dels paràmetres de les distribucions (maximitzar la coincidència amb la realitat de les distribucions donades les dades.

- Endevinem els paràmetres de les distribucions considerant que tenim només les probabilitats de pertànyer a cada clúster i no els clusters. Les probabilitats actuen com a pesos.

$$\mu_A = \frac{w_1 x_1 + w_2 x_2 + \dots w_n x_n}{w_1 + w_2 + \dots w_n}$$

$$\sigma_A^2 = \frac{w_1(x_1 - \mu)^2 + w_2(x_2 - \mu)^2 + \dots w_n(x_n - \mu)^2}{w_1 + w_2 + \dots w_n}$$

on wi es la probabilitat que l'objecte i pertanyi al cluster A i es sumen tots els objectes (no només els del A).

L'algoritme tendeix a convergir en un punt (mai arribar).

- Veiem com arriba a convergir calculant la versemblança de les dades amb els paràmetres, multiplicant les probabilitats dels objectes individuals i:

$$\prod_i (P_A P(x_i|A) + P_B P(x_i|B))$$

- Mentre creixi la mesura anterior i fins que sigui despreciable tenim que anar iterant (amb compte de que aquesta convergència pot ser un màxim local).

12.2. ASSOCIACIO

En poques paraules, les regles d'associació poden servir per a:
- Establir vincles (associacions) entre els registres
- trobar seqüències similars
- trobar patrons seqüencials

Els resultats (els vincles trobats entre els registres) a vegades s'utilitzen coma punt de partida quan no es coneixen exactament els patrons o preguntes a buscar.
Un altre punt a favor és que existeixen multitud de regles per a tractar tot tipus de dades.
Usualment els mètodes d'associació es fonamenten en tècniques estadístiques.
Per a obtenir patrons seqüencials necessitem relacionar cada succés a una dada o ordinal.

No totes les regles o combinacions de regles trobades seran d'utilitat. Els paràmetres per a mesurar la qualitat d'aquestes regles es:
- suport (support): nombre d'instàncies predites correctament

$\text{suport}(A \Rightarrow B) = P(A \cup B)$
- confiança (confidence): proporció de nombre d'instàncies a les quals s'aplica la regla

$\text{confiança}(A \Rightarrow B) = P(B|A) = \text{suport}(A \cup B) / \text{suport}(A)$

Per una altra banda:
- Lift: compara la probabilitat de trobar els productes relacionats en el resultat:

P(C y R)/P(C)P(R). Si es major que 1 llavors la regla és millor que en el cas aleatori. Hi ha que tenir en compte que pot produir error si el suport es baix.

Cada atribut-valor (ítem) volem que cobreixi moltes instàncies. Després transformarem els ítems amb prou confiança en regles. Cada ítem pot generar una regla o més d'una (o cap).
Això que acabo d'esmentar es exactament el que anuncia un dels algoritmes (amb les múltiples variants i extensions que té) més usats dins d'aquest tipus d'anàlisi.

Apriori:

Obtenir items-sets (conjunt de valors que es repeteixen) de un determinat tamany per a combinar-los en regles.
1r. Per cada conjunt l de ítems, generar tots els seus subconjunts
2n Si: suport(l)/suport(s)≥nivell de confiança; llavors per a cada subconjunt $s \subset l$, es genera una regla $s \Rightarrow (l-s)$.

Encara que és molt eficient per a grans volums de dades y alguns gestors de bases de dades inclòs poden fer-ho dins del propi programa gestor, hi ha per contra que per a algunes dades d'entrada els resultats intermitjos consumeixen gran quantitat de recursos.
Requisits:
- No necessita explicitar els atributs dels costats dret o esquerre de les regles, perquè es generen de manera automàtica.
- Hi han moltes varietats per a tractar tot tipus de dades

- S'ha d'especificar el mínim suport
- Y s'ha d'especificar el màxim nombre de regles.

A vegades quan parlem de regles d'associació podem veure útil (només en alguns casos) el pensar també en les negacions de les condicions. Encara que resulta traumàtic quan pensem en recursos consumits, a vegades negar alguns ítems es molt interessant.

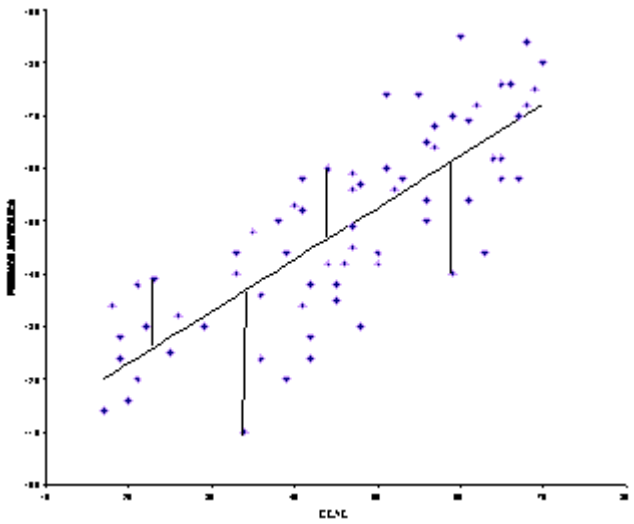
12.3. PREDICCIÓ

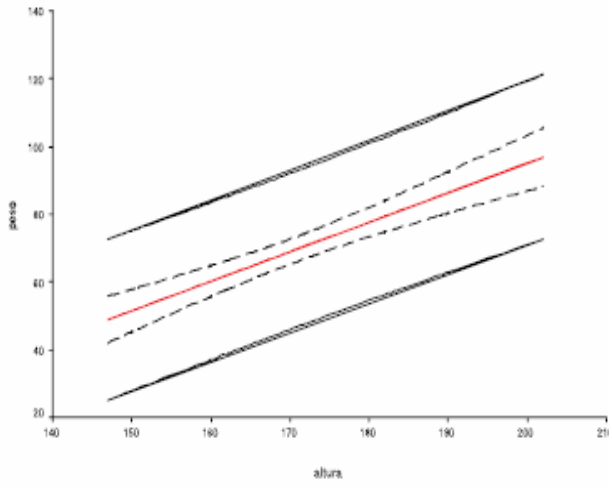
Estimació o regressió, introdueix un model per a predir el valor de la classe donats valors dels atributs. Treballarem amb funcions continues.
La diferència entre les nominacions regressió i predicció és:

- Predicció: es dona un conjunt de dades y un model que treballa sobre ells, es tracta de predir el valor d'un atribut específic que encara no es té. També tracta a vegades de validar hipòtesis que involucren altres dades.
- S'analitzen la dependència entre valors d'atributs. L'atribut dependent es pot predir aplicant el model i el valor dels atributs independents.

Els mètodes de predicció són estan molt interrelacionats amb altres mètodes. Per exemple, molts treballs de data Mining mostren la classificació i la estimació com a subnivells de la predicció. D'altres treballen molt les diferències entre classificació i predicció.
Com aquest primer, la predicció (i sobretot les regressions) s'usen com a feines prèvies per a tenir una idea global de les dades que estem manejant.
Sempre hem de tenir en compte que en aquests mètodes usem variables contínues (el tret diferenciador dels arbres de predicció).
Els seus algoritmes són usats com a previs d'altres models, ja que trobar la relació existent entre dos varibles sol ser un dels pasos inicials (sobretot en la seva forma més simple), per tant veurem més a fons les tècniques en altres apartats (estadística, probabilitats, arbres…).
Pex. Els arbres de regressió són similars als arbres de desició però basats en tècniques estadístiques.

El principal mètode de la predicció és la regressió (lineal i múltiple i no lineal). Encara que ja he deixat clar que no són el mateix, un cop feta una regressió i trobades les dependències es pot predir el comportament de nous valors no analitzats encara:





Per una altra banda, els arbres de predicció són els que usarem un cop conformats per a "deduir" el comportament de noves dades.

12.4. CLASSIFICACIÓ

Donat un conjunt de dades, volem determinar a quina classe pertany cada ítem. Les classes o grups són excloents entre sí. És pot veure com l'esclariment d'una dependència, en la que l'atribut dependent pot prendre un valor entre vàries classes, ja conegudes.

Sent la manera de analitzar més antiga i utilitzada té a més nombroses tècniques:
Els requisits són:

- Donar l'(els) atribut(s) de desició o classe (els usats per a construir la classe d'equivalència en els mètodes supervisats, una per cada valor o combinació de valors dels atributs) i els atributs de condicions (els usats per a descriure mitjançant el procés d'inducció les classes d'equivalència)
- Hi han algunes tècniques que només tracten dades numèriques
- Nombre màxim de precondicions
- Suport mínim de les regles

Està clar que el conjunt de dades d'entrenament i el conjunt de dades de prova han de ser disjunts.

ARBRES DE DESICIÓ

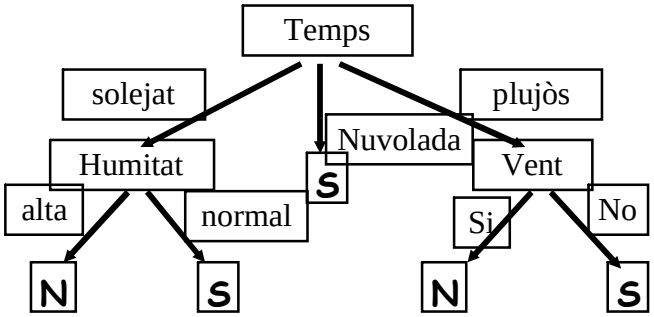
- Descobriment de regles i relacions
- Subdivisió successiva del conjunt de dades
- Molt útils quan es vol fer classificacions molt amplies o prediccions
- Millor variables que es puguin dividir en pocs valors

Primer es crea un arbre amb unes dades de mostra:

1. Els nodes representen la verificació d'una condició sobre un atribut
2. Les branques representen el valor d'una condició
3. Les fulles representen les etiquetes de classe

Ambient	Temperatura	Humitat	Vent	Classe
Solejat	**Alta**	**alta**	**No**	**N**
Solejat	**Alta**	**alta**	**Si**	**N**
Nuvolada	**Alta**	**alta**	**No**	**S**
Plujós	**Mitja**	**alta**	**No**	**S**
Plujós	**Baixa**	**normal**	**No**	**S**
Plujós	**Baga**	**normal**	**Si**	**N**
Nuvolada	**Baixa**	**normal**	**Si**	**S**
Solejat	**Mitja**	**alta**	**No**	**N**
Solejat	**Baixa**	**normal**	**No**	**S**

Plujós	**Mitja**	**normal**	**No**	**S**
Solejat	**Mitja**	**normal**	**Si**	**S**
Nuvolada	**Mitja**	**alta**	**Si**	**S**
Nuvolada	**Alta**	**normal**	**No**	**S**
Plujós	**Mitja**	**alta**	**Si**	**N**



Ara aquest arbre el podem utilitzar per a classificar una mostra de la quan no en sabem les classes: mirem els valors dels seus atributs i anem baixant per les rames.
Realment el que hem fet es posar les regles en forma de arbre.

El arbre resultant pot ser, en les bases de dades més complicades, molt complex. Llavors es fa necessari un **mètode de poda,** per a que l'arbre no tingui més nivells ni particions dels necessaris per a aconseguir una predicció.
Encara que hi ha diverses tècniques, un procediment podria ser:

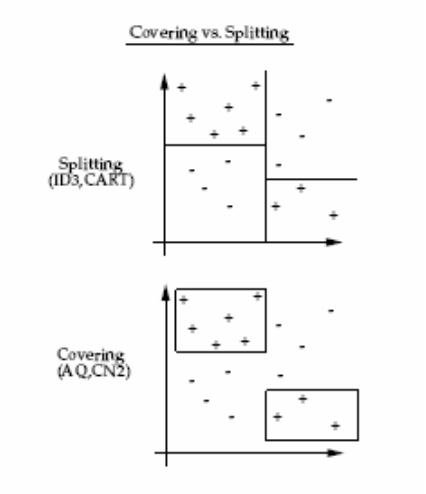
Procedimiento *poda*
(1) Sea N la raíz de Arbol
(2) Calcula el error E de N
(3) **si** N es una hoja, **entonces** Arbol $= N$, Error $= E$
(4) **para cada** subárbol con raíz en N,
aplica *poda* para obtener un árbol podado y una estimación de error
Calcula el error esperado Es si dividimos en N y usamos los subárboles para clasificar
(5) **si** $Es > E$, Arbol $= N$ y Error $= E$.
sino Arbol = árbol con raíz N y subárboles (los árboles podados en el paso (4), y Error $= Es$.

INDUCCIÓ DE REGLES DE CLASSIFICACIÓ

La inducció es raonar les propietats d'individus a propietats de conjunts d'individus. A partir de esta definició, induir les regles serà el mateix aplicat al camp que ara ens ocupa – per a generar i contrastar arbres de decisió o regles y patrons a partir de les dades d'entrada-, i es fa de dues maneres:

- afegint atributs a l'arbre que s'està construint maximitzant la separació entre classes (basat en l'splitting: dividir el conjunt de dades en subconjunts considerant un atribut seleccionat per una heurística particular)

- afegint proves a cada regla que s'està construint maximitzant la cobertura minimitzant errors (basat en el covering: trobar condicions de regles (atribut-valor) que cobreixi la major quantitat d'exemples d'una classe i la menor quantitat de la resta de classes.



BAYESIANA

Aquestes tècniques permeten aprendre sobre les relacions de dependència i casualitat, combinar els coneixements de les dades, evitant el sobreajust de les dades y permetent treballar amb bases de dades incompletes.

Bases:

Donat un conjunt de dades, la probabilitat a posteriori de una hipòtesi és:

$$P(w_i / x) = \frac{p(x / w_i)P(w_i)}{P(x)} \qquad P(x) = \sum_i^J p(x / w_i)P(w_i)$$

Per tant requereix un coneixement previ de les probabilitats.

Probabilitats prèvies:

$P(A_i)= (a_i+1)/(s+1)$ si i=k

$P(A_i)= a_i/(s+1)$ si i≠k

En un classificador de Bayes:

- Un error de classificació es pot quantificar com un cost
- L'esperança de cost (el risc) es pot acceptar com a criteri d'optimització.

Es el classificador que dona menys error.

Funció de cost:

$$L_{j\hat{k}} = L(C_j \mid \hat{C}_k) : \Omega \times \Omega \to \Re^+$$

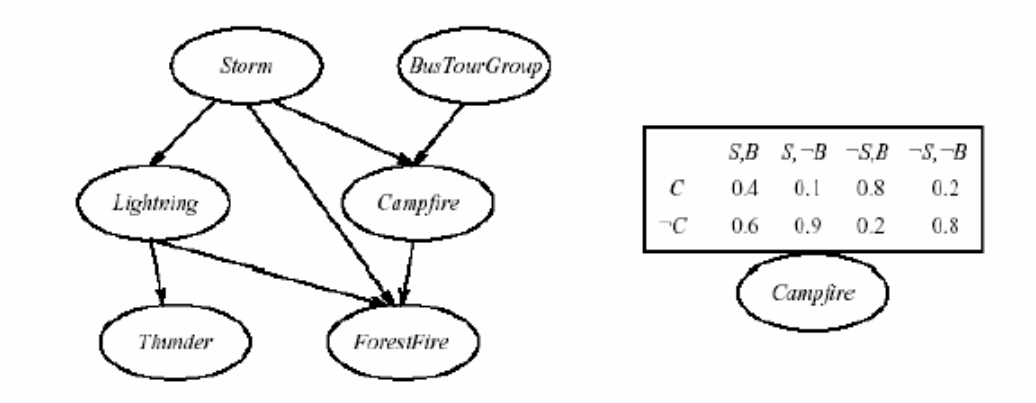
Risc per a una classe:

$$R_k = R(\hat{C}_k \mid x \in C_k) = E\left\{L_{j\hat{k}} \mid x \in C_k\right\} = \sum_{j=1}^{N} L_{j\hat{k}} \int_{\Re_j} p_{x|C_k}(x \mid C_k)dx$$

Risc mitjà del classificador:

$$R = \sum_{k=1}^{N} R_k \Pr(C_k) = \sum_{j=1}^{N} \int_{\mathcal{R}_j} \left\{ \sum_{k=1}^{N} L_{jk} p_{x|C_k}(x \mid C_k) \Pr(C_k) \right\} dx$$

Quan ho presentem en mode de graf cada node (que te associada una taula de probabilitat condicional) representa una variable aleatòria i cada arc una dependència.

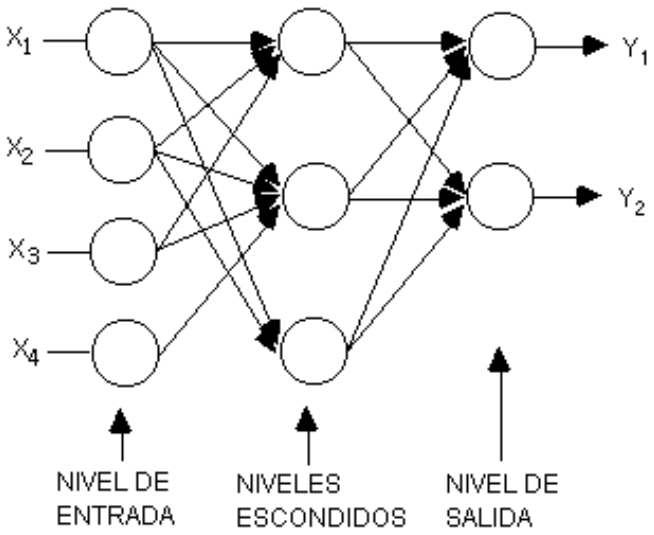


XARXES NEURONALS

Una xarxa neuronal es un model computacional amb un conjunt de propietats específiques, com són l'habilitat d'adaptar-se o aprendre, generalitzar o organitzar la informació, tot això basat en un procés eminentment paral·lel. (Kröse y van der Smagt,1993)

A mode de connexions neuronals d'un cervell (d'aquí el nom), permet construir un model de comportament a partir d'una determinada quantitat d'exemples, trets d'una quantitat de variables descriptives del comportament. La xarxa va aprenent del exemples per a acabar transformant-se en un model susceptible de poder explicar el comportament observat en funció de les variables descriptives.
Al ser un model automàtic i directe des de les dades no necessita gran quantitat de recursos intermitjos.
La xarxa simula grups de neurones, que estén relacionades. Les dades s'introdueixen en la primera capa, la capa d'entrades, i cada capa va transferint la informació a les veïnes (que poden tenir un pes específic per als valors, a mode de pes-importància-connexió de les sinapsis). Quan les dades arriben a l'última capa, la de sortida, el valor resultant es el pres com a resultat de la red.

Lo més difícil i la gran desavantatge en aquesta tècnica es entendre el resultat, ja que el pas per diferents capes, amb diferents pesos pren rellevància sobre les dades. A més l'aprenentatge té una duració que a priori no coneixem, i que durant la cual es van modificant les connexions i els pesos d'aquestes fins que aquestes romanen estables.



LÒGICA BORROSA

La lògica borrosa el que fa realment és definir modes de comportament més aproximats que exactes, cosa de la que difereix de la lògica clàssica.
Exactament el que tracta es d'apropar el llenguatge matemàtic al llenguatge humà, molt més imprecís. La gran difusió que estan tenint sistemes basats en aquesta lògica (Fuzzy Logic) estan tornant-la important en moltes àrees de treball. Exemple: "Si fa molta calor i no hi ha moltes fonts llavors es vendran molts gelats"
Per tant aquí quan tractem de posar un individu a un grup podem trobar-nos en que aquest es pugui trobar en més d'un, a causa de les difuses barreres que hi ha en aquests, i que sigui la situació exacta qui determini on pertany.
Les corbes de possibilitat modelitzen matemàticament aquest fet

TÈCNIQUES GENÈTIQUES

Són tècniques d'optimització que usen processos com les combinacions genètiques, mutacions i selecció natural en un disseny basat en els conceptes d'evolució.Com passa amb la lògica borrosa, és la falta de barreres, de determinació, la que converteix els resultats en difícils a l'hora de treure models fiables.

- Mòdul evolutiu: mecanisme d'interpretació de la informació (decodificació) y funció d'avaluació (mesura la qualitat). Només aquí existeix la informació del domini
- Mòdul Poblacional: té una representació poblacional i tècniques per a manipular-la. Aquí es defineix el tamany de la població i la condició de determinació.
- Mòdul Reproductiu: conté els operadors genètics.

Aquest algoritme permet obtenir solucions a un problema que no té cap mètode de resolució descrit de forma precisa o quan la solució exacta, si es coneguda, es massa complicada per a trobar-la en un temps real. Nombroses solucions seran creades a l'atzar, segons una forma definida.

13. SYNERA

"La principal diferencia entre este sistema de Business Intelligence es que sabe qué información se ha de buscar. Se trata de un análisis dinámico ilimitado", Tony Barbosa, director general de la compañía Synera Systems.

13.1. Introducció

- És un sistema avançat i intel·ligent d'interpretació y exploració de dades. Interactiu, potent i escalable.
- Solució orientada a realitzar anàlisis en temps real sobre les dades emmagatzemades, descobrint de manera natural les regles i pautes de tota una estructura.
- Synera Explorer permet consultar qualsevol informació que pugui contenir la base de coneixement. Fa possible l'execució de consultes iteratives i sense estructura, basades en l'exploració visual i l'anàlisi. L'usuari pot fer consultes multidimensionals sobre totes les dades, generar expressions o segmentacions en funció de valors estadístics.
- Synera Discovery, eina d'anàlisi, incorpora mètodes de Mineria de Dades i detecció de patrons que podrien passar desapercebuts.

13.2. Paquet Synera. Esquema del programari

*"**synera** proporciona el Knowledge Discovery y la administración directa del Data Warehouse."*

- Synera es un software especialitzat en gestionar l'accés a les bases de coneixement des d'aplicacions client, que es connecten mitjançant protocol TCP/IP-> per tant ens servirà per a poder treballar molt agust des de la intranet de l'empresa Motorpress, que suporta aquest protocol.
- Administra les connexions dels usuaris i els seus privilegis d'accés a la base del coneixement, planificant processos i serveis asíncrons ->podrem gestionar l'accés a les dades des dels terminals dels diferents departaments.
- Guarda les dades com a valors i no com a registres, evitant així la repetició de dades->optimitza espai.
- Gestiona les consultes i accepta funcions de àlgebra de conjunts permetent seleccionar subconjunts d'informació en funció de criteris de búsqueda incrementals i dinàmics, i també de diferents parts de la base de coneixement. A més diferents usuaris poden tenir consultes parcials que després un altre usarà->molt útil quan es tinguin diferents blocs de dades i per a quan es treballi asíncronament des de diferents terminals.
- El resultat de una consulta es pot afegir com a informació nova de la base de coneixement.
- Usant tècniques de Data Mining permet detectar, asíncrona i independentment, patrons de comportament en les dades.

Synera Engine: arranc i parada del servidor Synera.
Synera Loader: carregar grans volums de dades.
Synera Users: definir i gestionar els usuaris de Synera, els seus permisos i privilegis.
Synera Configurator: crear i configurar una base de coneixement Synera.
Synera Explorer: Emmagatzemar i analitzar dades usant Synera (consultes, us de sentències SQL, editar i analitzar dades...)
Synera Discovery: analitzar les dades usant tècniques de Mineria de Dades: MBA i Clustering.

13.3. Presentació i selecció de dades

Les dades s'han aconseguit mitjançant uns qüestionaris emplenats tant des de la web de Maxi Tuning com en concentracions. També disposem d'unes dades recollides des d'un formulari de la revista, però amb la intenció de conèixer el perfil del potencials compradors, més que els que ja la llegeixen, començaré amb les dades primeres. A més les preguntes fetes des del formulari de la revista estan incloses en les del segon formulari, per tant ens poden ser útils després per a veure si el que hem trobat és correspon amb els lectors de la revista (si més no amb els que han contestat).

El recopilament informàtic es va realitzar mitjançant taules d'accés relacionades, tenint la enquesta (totes les relacions) com a principal.
A aquests registres vaig incloure els que hem van arribar en paperetes.

Id	Enter	Automàtic	Identitat del registre (i explicació)
Nombre	Char	Text de lliure introducció	Nom/Sobrenom
E-Mail	Char	Text de lliure introducció	Direcció e-mail
Edad	Enter	Text de lliure introducció	Edat
Sexo	Char	Elecció: Hombre Mujer	Sexe
Hermanos	Booleà	Elecció (validació) Sí/No	Tens germans?
Provincia	Char	Elecció de la comunitat	Provincia
Habitantes	Char	Elecció: - 1,500 - 10,000 - 50,000 -100,000 -250,000 -500,000 +500,000	Habitats de la teva ciutat
Estudias	Booleà	Sí/No	Estudies?
Trabajas	Booleà	Sí/No	Treballes?

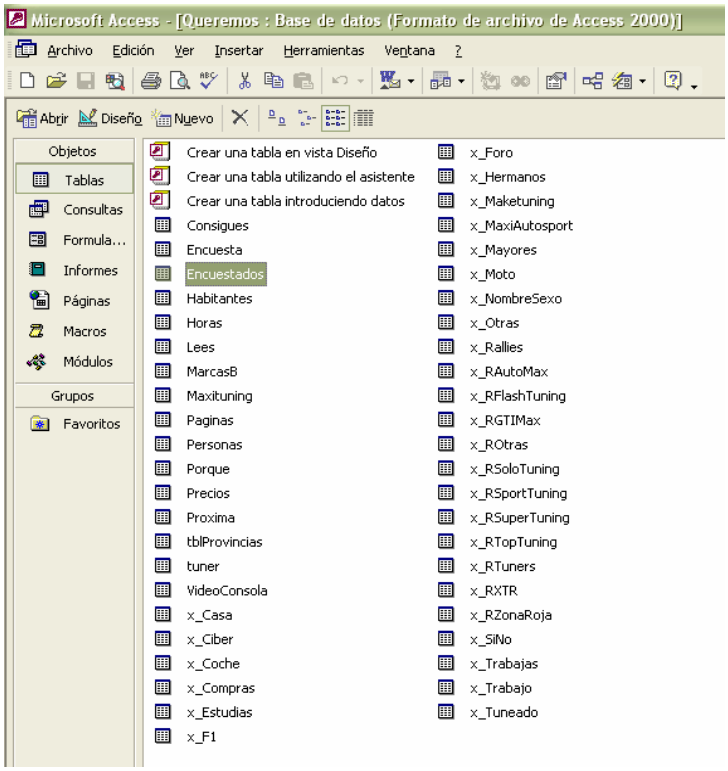
VideoConsola	Char	Elecció: PlayStation Xbox GameCube Ninguna	
Casa	Booleà	Sí/No	Et connectes a internet desde casa?
Cibercafé	Booleà	Sí/No	Et conectes a internet des d'un ciber?
Trabajo	Booleà	Sí/No	Et connectes a internet des del treball?
ComprasInternet	Booleà	Sí/No	Compres per Internet?
Horas	Char	Elecció: 2 4 6 8 +10	Quantes hores passes navegant a la setmana?
Foro	Char	Sí/No	Uses el foro de la web de Maxi?
Coche	Booleà	Sí/No	Tens Cotxe?
Marca	Char	Elecció: Ford Seat Opel Renault Citroën Alfa Romeo Audi BMW Chrysler Daewoo Ferrari Fiat Honda Hyundai Kia Lada Lancia Land Rover Lexus Lotus Mazda Mercedes MG Mini Cooper Mitsubishi Nissan Peugeot Porsche Piaggio Rover Saab Subaru	Marca del Cotxe

		Smart Skoda Suzuki Toyota Volvo VolksWagen Yamaha Desconocida Talbot	
Modelo	Char	Text de lliure introducció	Model del cotxe
Tuneado	Booleà	Sí/No	Lo tienes tuneado?
Tuner	Char	Elecció: Barroco Fino Import Alemán Street Otros	Estil que li has aplicat
Inviertes	Char	-200€ de 200 a 600€ de 600 a 1,000€ de 1,000 a 2,000€ de 2,000 a 4,000€ + 4,000€	Quan gastes al més en tuning?
GastosProximo	Char	Elecció 1 mes 2 meses 3 meses 6 meses Próximo año	Quant penses tornar a invertir en tuning?
Gastos	Char	-200€ de 200 a 600€ de 600 a 1,000€ de 1,000 a 2,000€ de 2,000 a 4,000€ + 4,000€	Gastos en hobbys al mes?
Consigues	Char	Elecció: Ahorrando Préstamo Regalo Otros	Com aconsegueixes els diners?
Moto	Booleà	Sí/No	Tens moto?
Maxituning	Char	Elecció: Desde el número 1 Desde hace 2 años Desde hace 4 años Desde hace 2 meses	Des de quan llegeixes MaxiTuning?
Suscrito	Booleà	Elecció Si/No	Estàs subscrit?
Porque_no_	Char	Estoy suscrito a demasiadas No lees cada mes la revista	Per què no ho estàs?

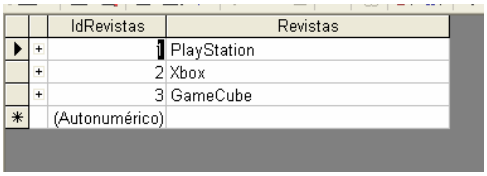
		El precio de abono es muy alto Otros	
Porque_Otros	Char	(Inicialment lliure introducció de text) A veces la compra mi hermano/a A veces la compra algún amigo/a Prefiero comprarla en el kiosco mensualmente Tengo un kiosco No sé cómo subscribirme A mi novia no le gusta la revista y no me deja Hay meses que no tengo el dinero A veces se me olvida comprarla Acabo de descubrirla Acabo de meterme en el mundo del tuning No puedo pagarlo todo de golpe Tengo otros gastos Estoy pensando en suscribirme Voy a suscribirme en breve Cada mes la compra un colega Mis padres no me dejan Compro demasiadas revistas Compro los números que me llaman la atención Compro otro tipo de revista Compro varias revistas al mes No me fío que me llegue Tarda mucho en llegar No me gusta su contenido En donde vivo no la venden siempre En el kiosco me la reservan todos los meses Porque es mi hobby Lo está mi hermano/a Estoy suscrito a otra/s Estuve suscrito y prefiero comprarla yo No compro la revista todos los meses No tengo tiempo de comprarla La compartimos entre los compañeros Apenas tengo tiempo para leer No veo las ventajas por el precio La leo en algún bar/tienda Me la deja un amigo/a La veo en internet No he pensado nunca en	Per què otros?

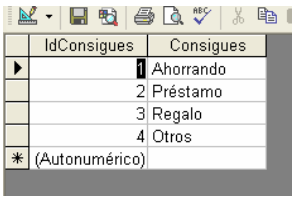
		suscribirme Lo envié pero no me suscribieron No lo sé Porque me gusta vuestra revista Porque la compro todos los meses Por mi trabajo No me interesa Soy de fuera y es difícil conseguir aquí	
Cuantos_leen_tu_MT	Char	Elecció: 1 2 4 +4	Quantes persones llegeixen la teva revista?
Comprasotras	Booleà	Sí/No	Compres unes altres revistes?
Tuners	Booleà	Sí/No	
AutoMax	Booleà	Sí/No	
FlashTuning	Booleà	Sí/No	
GTIMax	Booleà	Sí/No	
SoloTuning	Booleà	Sí/No	
SportTuning	Booleà	Sí/No	
XTR	Booleà	Sí/No	
TopTuning	Booleà	Sí/No	
SuperTuning	Booleà	Sí/No	
Otras	Booleà	Sí/No	
Autosport	Booleà	Sí/No	Conoces MaxiAutosport?
F1	Booleà	Sí/No	Te interesa la F1?
Rallyes	Booleà	Sí/No	Te interesan los Rallyes?
Maketas	Booleà	Sí/No	Te interesa el Maketuning?

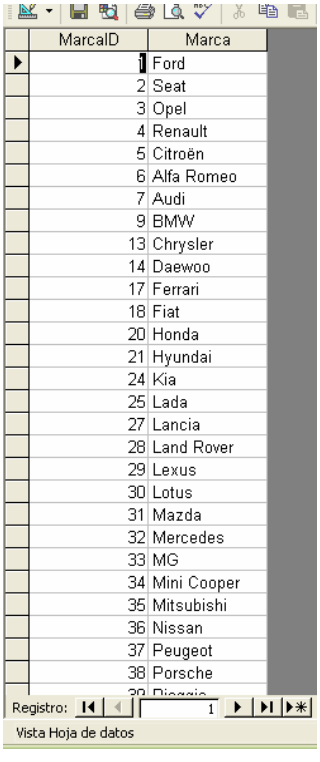
Nota: Els atributs queden així un cop realitzades les passes de preparació de dades, excepte el tipus que passo a explicar ara:

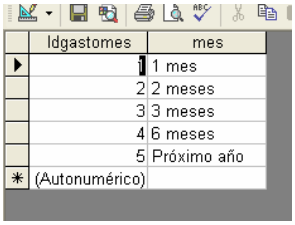


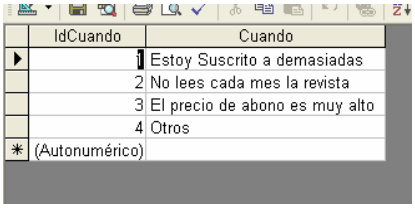
La majoria d'atributs són de selecció d'opcions d'una llista desplegable o de validació (si/no), d'aquí les taules dels valors de les seleccions; com per exemple:

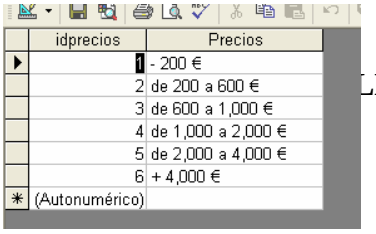












Tots els valors son doncs char.

Els de selecció si/no (marcar casella) es codifiquen internament amb 0/1 (compte, també char).
Hi ha que tenir també compte que al carregar-se la base de dades Enquesta hi ha alguns atributs (els que tenen menys valors diferents) que passen a començar des de 0, i no es corresponen amb els valors que aquí veiem.

El camp porque_otros no és d'elecció, però un cop netejat queda així.

Però al treballar amb el Synera canvio alguns dels valors enters "codificats" pel text per a poder usar-ho amb l'Explorer amb menys problemes., ja que l'accés molts de tipus char (els de selecció) que realment eren enters que servien com a índex del valor reals en la taula relacionada amb l'atribut X no l'aconseguia fer, ni reconèixer ni tractar. Aquesta tasca (canviar els valors) la faré amb el Synera Explorer.

SELECCIÓ DE DADES
Com que en un principi no se el que hem trobaré les carrego totes. Encara que alguns camps com l'e-mail EN AQUEST TREBALL* podrien excloure's (igual que, com després veure'm, molts, ja que no aporten informació que puguem usar per a l'objectiu de la tasca coneixement).

*Remarco que en aquest treball perquè per a un possible treball de text mining, per exemple, podríem estudiar els noms d'usuari etc.

13.4. Neteja de dades

Per a netejar les dades procedeixo a introduir les dades de Excel a una base de dades Acces, que m'ajuda a analitzar els punts que he explicat al punt 11.2. Neteja de dades :

- duplicitat de dades . Busco duplicitats en les dades, sobretot en les files, ja que les columnes són les mateixes en ambdues fonts de dades d'on provenen. En les entrades hem fixo en el mail (i miro si per a un mateix mail hi ha diferents persones o es la mateixa) i també miro si algun nom complet (nom+cognoms) es repeteix, comprovant que la població d'on ve també sigui la mateixa.
- sincronització de les referències a objectes o categories: Junt a amb l'anterior aquesta una de les parts més àrdues de treball es quan realitzo l'agrupament del text Porque_otros_ en categories que agrupen totes les coses que s'han introduït, ja que la lliure introducció de text havia fet diferents valors que volien dir el mateix com:

"La kiero comprar en el kiosco"
"Prefiero ir al kiosco"
"Porque la compro en el kiosco"... etc.

Les 2906 diferents respostes (després de la neteja) acaben en 47 només.

- dades incompletes : els camps buits de els completaré així:
    - E-Mail: Introduint "Ninguno"
    - Nombre: Introduint "Chico" o "Chica"
    - Edad: Com que l'edat majoritaria dels registres (429) és 18 m'aventuro a emplenar-lo amb aquest valor.
    - Videoconsola: Afegeixo una nova opció "Ninguna" i la selecciono.
    - Marca, Modelo, Tuner, Cuanto inviertes, Gastos en Hobbys, Desde cuando MT, Porque, Porque otros, Personas: com que van relacionats amb boleans anteriors ho deixo com està. Després comprovo que Synera hem respon.
- dades redundants : tindria com a redundant l'atribut Comprasotras, per què si hi ha alguna revista seleccionada ja ho sabriem. Però per a despres treballar amb Synera es interessant mantenir-la.
- dades incorrectes o inconsistents: els camps de lliure introducció de text (mail, nombre, edad, modelo –de coche-, por qué otros –de subscripción-) poden contenir dades incorrectes, sobretot en el format del mail. He actuat així:
    - mail: els que no contenen el format x@x.x els he eliminat.
    - nom: donat que hi ha la possible introducció de nick i la no obligatorietat d'introduir nom i cognoms, he respectat les dades inicials.
    - edat: només he mantingut els valors del rang 2-100; la resta els he eliminat per incorrectes (encara que hem queda un únic valor de 1 any que posaria la ma al foc de que no es real el deixo davant la possibilitat de que sigui el pare de un dels 6 subscriptors nounats que té la revista).
    - modelo: per la dificultat d'aconseguir una llista amb tots els models de cotxe existents, he repassat per si hi hagués algun error d'escriptura dels més corrents, però la resta la he respectat. Si hi ha algun model emplenat però no hi ha marca, he posat la marca si la sabia i si no he posat com a marca el valor "desconegut".

    - por qué otros: degut al treball anterior descrit en el primer punt no he necessitat treballar sobre aquest valor.
- dades envellides : encara que de moment podem confiar amb la fiabilitat de les dades, hi ha que tenir molt en compte que aquesta base de dades es quedarà obsoleta molt prompte, a causa de camps com l'edat (que podria solucionar-se canviant l'entrada de l'edat per la data de naixement) o la situació personal determinada en un moment donat, quan s'emplenen els formularis (si es té cotxe, les inversions, quines revistes es compren…)
- dades esbiaixades : està clar que al treballar amb clients "potencials" (gent que s'ha passejat per alguna concentració, que han passat per la web…) ja estem treballant amb una franja de persones molt concreta. Però per a la nostra tasca (el trobar els patrons dels potencials compradors de la revista) ja ens va molt bé.

13.5. Transformació de les dades

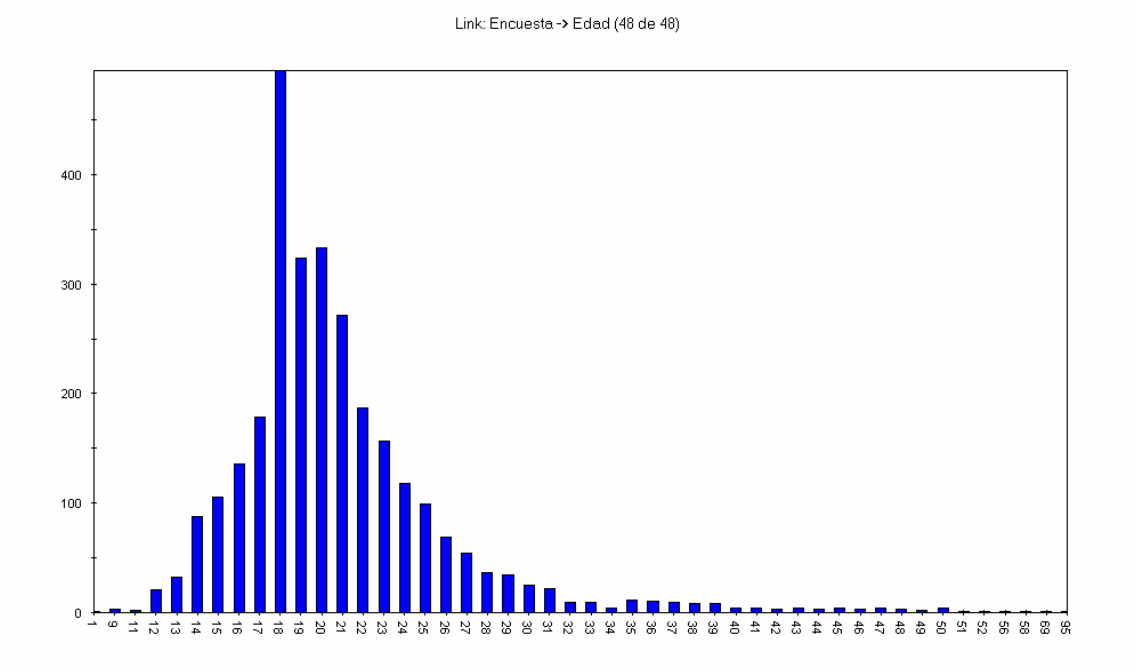
Pel que fa a la transformació de dades, en un primer moment no he tocat res més després de la neteja, ja que el que la major part d'atributs (sobretot numèrics, com els rangs de despeses) fossin d'elecció i no de lliure introducció fan que les dades estiguin manejables, a més d'estalviar que surtin dades esbiaixades. Després per a intentar trobar patrons que se m' escapin ja jugaré amb aquestes dades.

13.6. Tasques prèvies a l'anàlisi. Coneixement inicial.

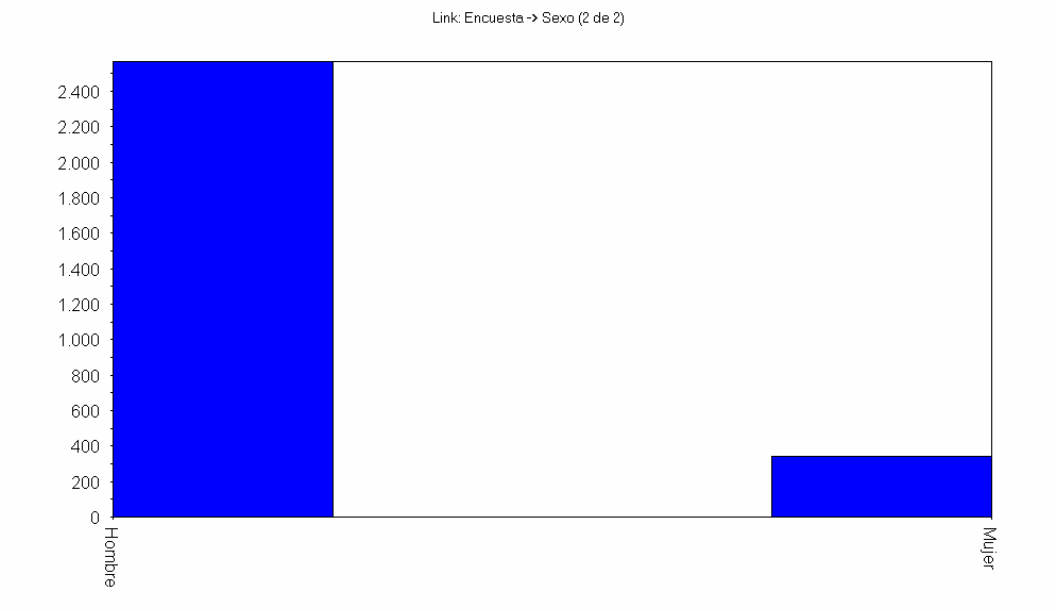
Puc extreure gran quantitat d'informació inicial a partir de tractar els atributs inicialment. Usant la opció de "Ver valores" des del diagrama d'ítems i les estadístiques puc extreure les primeres apreciacions (quan adjunto més d'un gràfic es fruit d'un residu de treball inicial en word, però que considero que completa la informació visual):

Edad

Com veig les edats a partir dels 30 anys no són rellevants, així com per sota dels 12.
La mitjana son 20,80 anys.

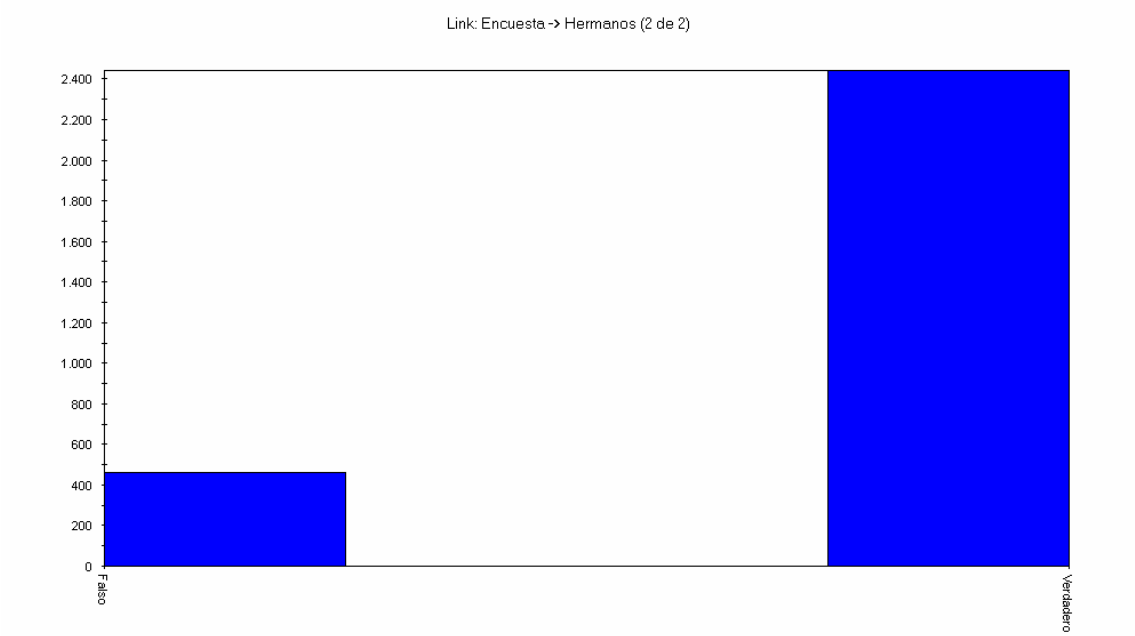


Sexo



De manera contundent són homes els interessats pel tuning.

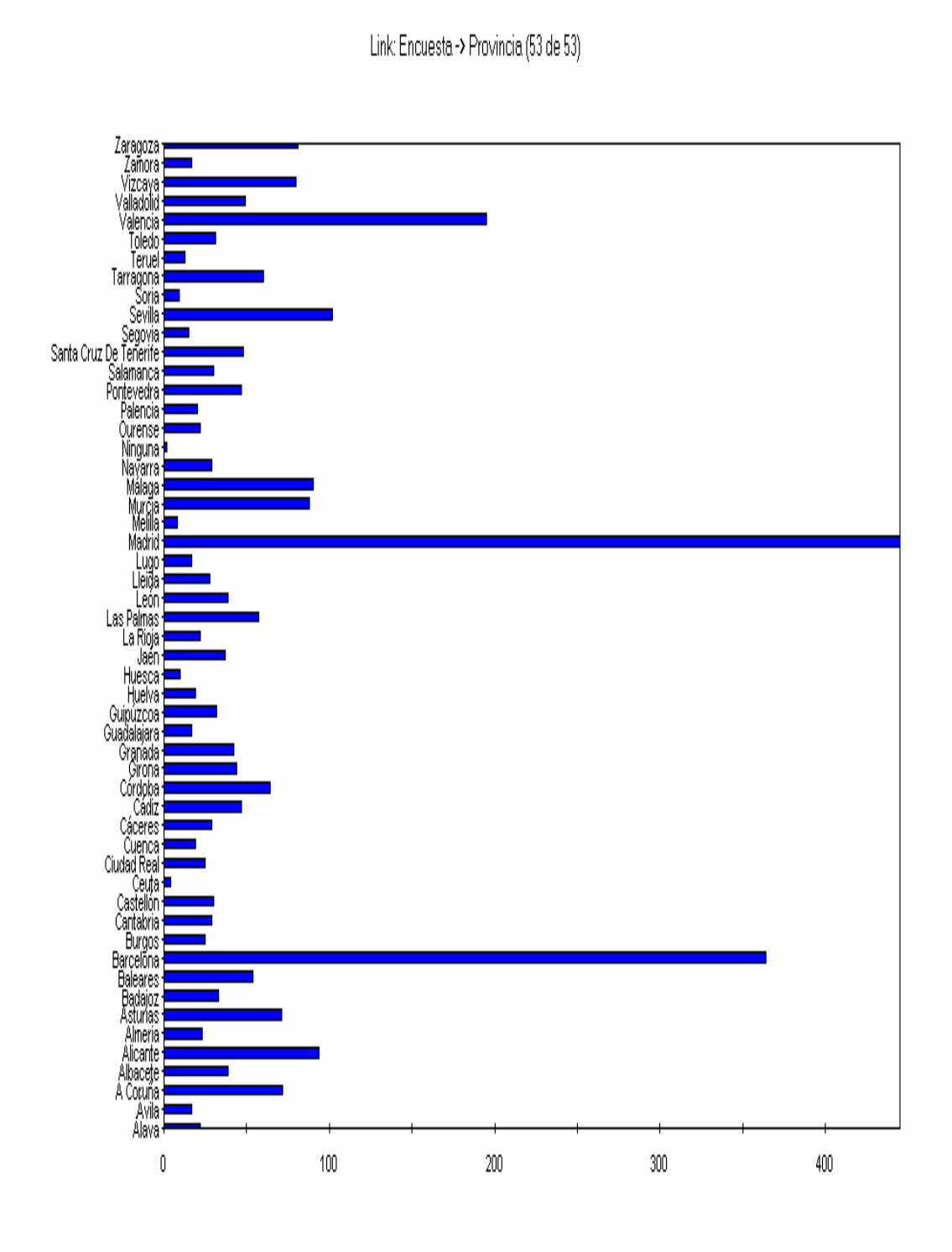
Hermanos



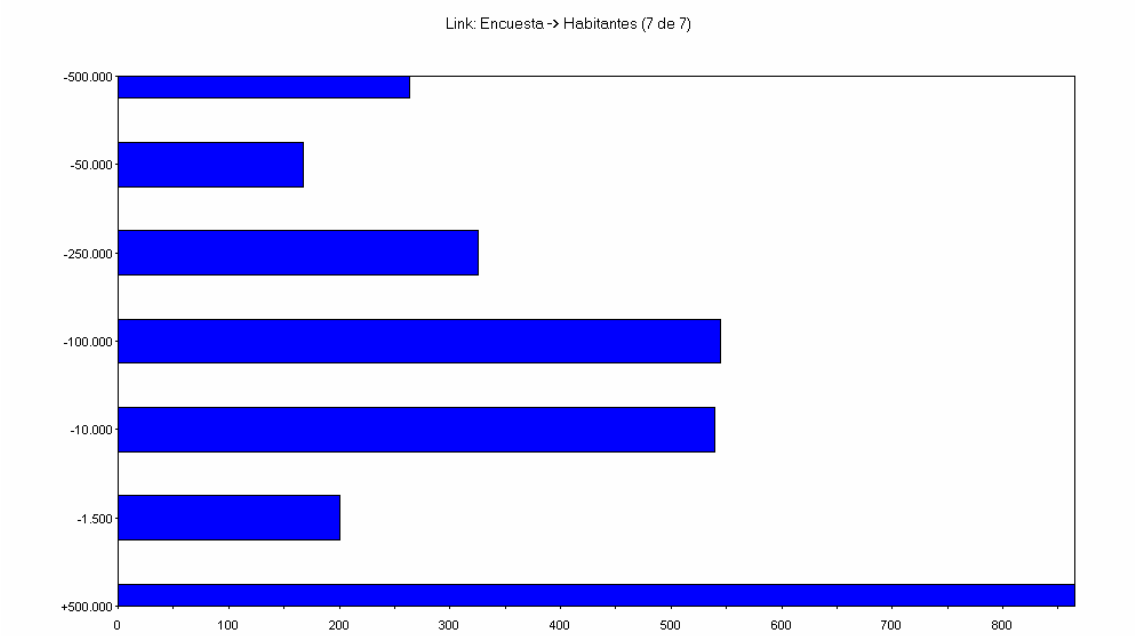
Aquest atribut no donaria informació rellevant, ja que la proporció es similar a la de la que es troba en la població total.

Provincia

Madrid, Barcelona i València són (com era d'esperar) on es concentren el major nombre de registres. Haurà que estudiar-se la relació de les províncies amb els altres ítems. Per contra hi ha altres províncies quasi inexistents en quan a enquestats.

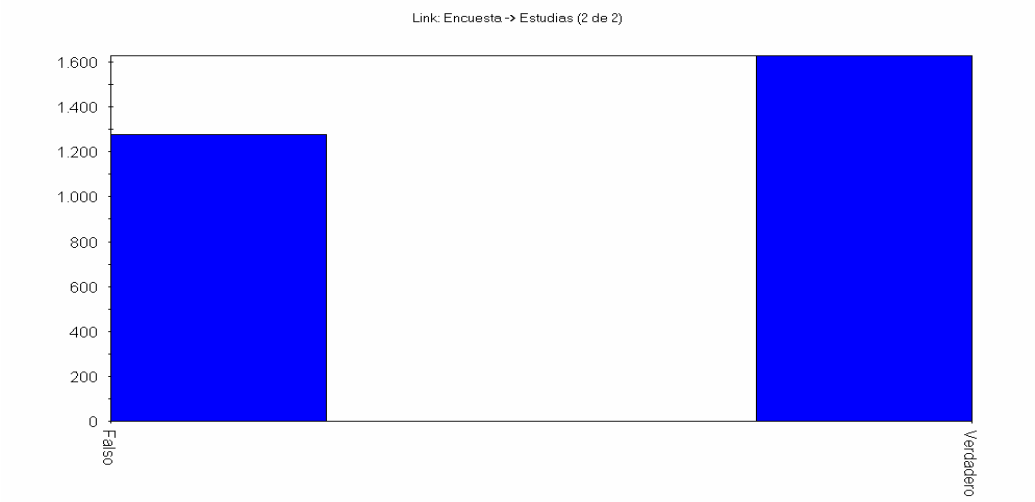

Link: Encuesta -> Provincia (53 de 53)
Zaragoza
Zamora
Vizcaya
Valladolid
Valencia
Toledo
Teruel
Tarragona
Soria
Sevilla
Segovia
Santa Cruz De Tenerife
Salamanca
Pontevedra
Palencia
Ourense
Ninguna
Navarra
Málaga
Murcia
Melilla
Madrid
Lugo
Lleida
León
Las Palmas
La Rioja
Jaén
Huesca
Huelva
Guipuzcoa
Guadalajara
Granada
Girona
Córdoba
Cádiz
Cáceres
Cuenca
Ciudad Real
Ceuta
Castellón
Cantabria
Burgos
Barcelona
Baleares
Badajoz
Asturias
Almería
Alicante
Albacete
A Coruña
Avila
Alava
0
100
200
300
400

Habitantes

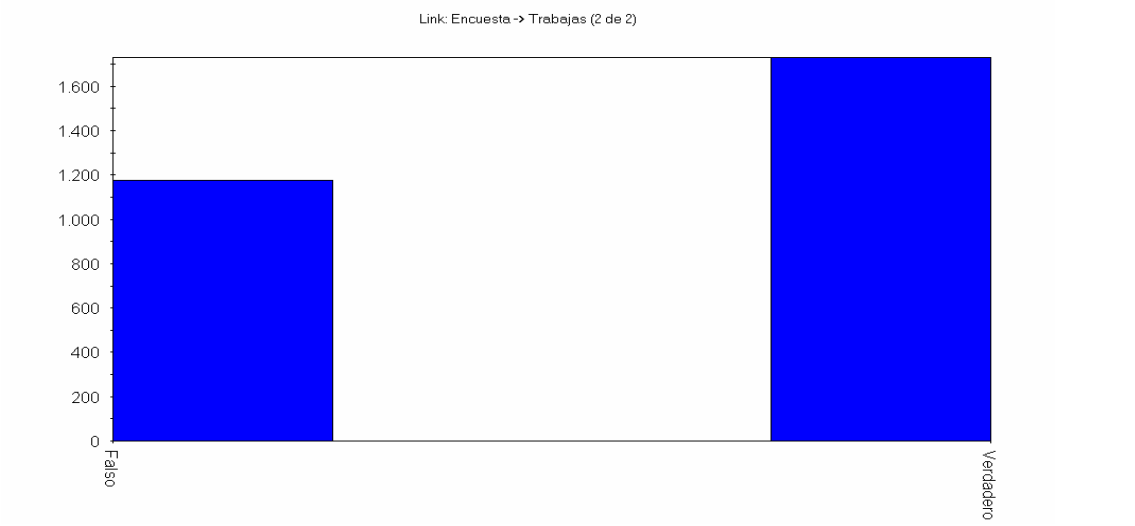


La distribució es estranya. Quasi tres quarts dels enquestats serien de ciutats de més de 50.000 habitants. Per tant pareix que és més popular el tuning a les ciutats

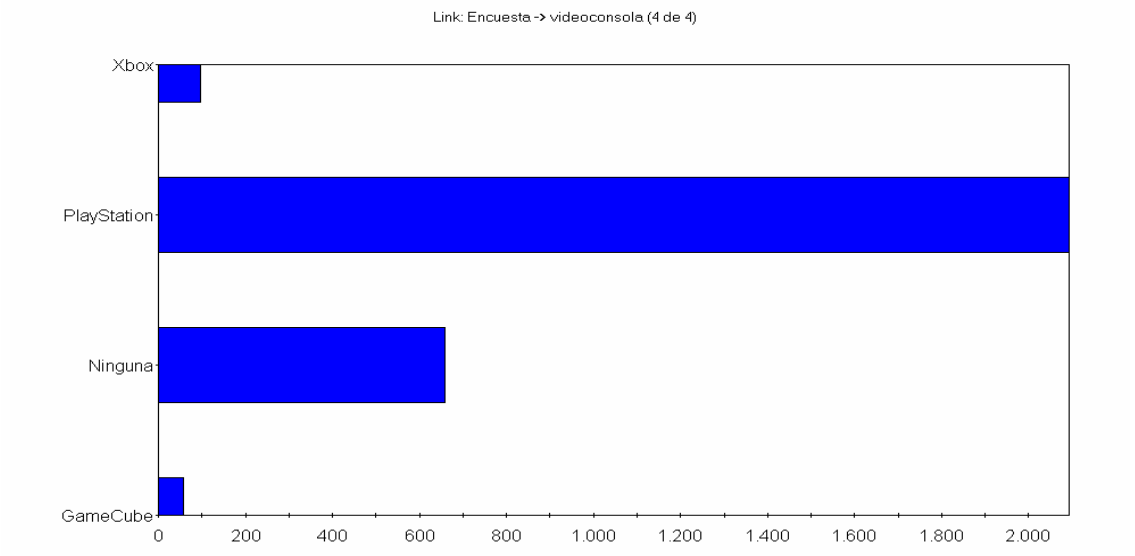
Estudiantes

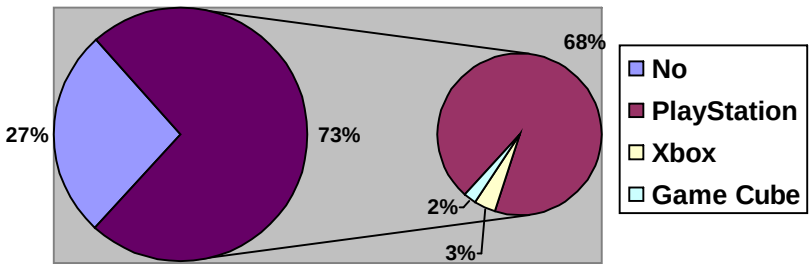


Trabajadores



VideoConsola

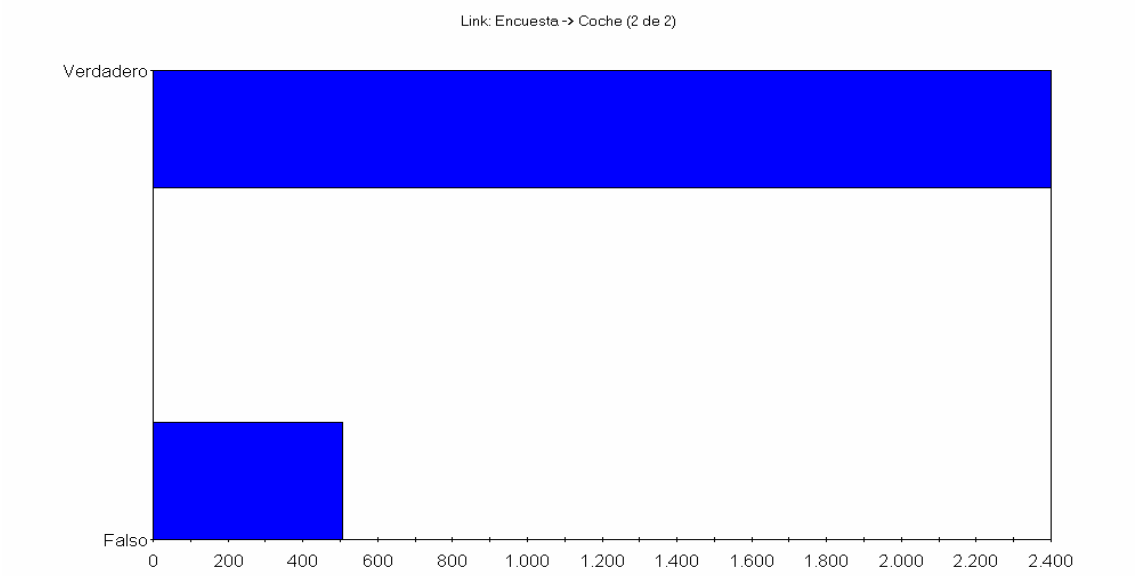


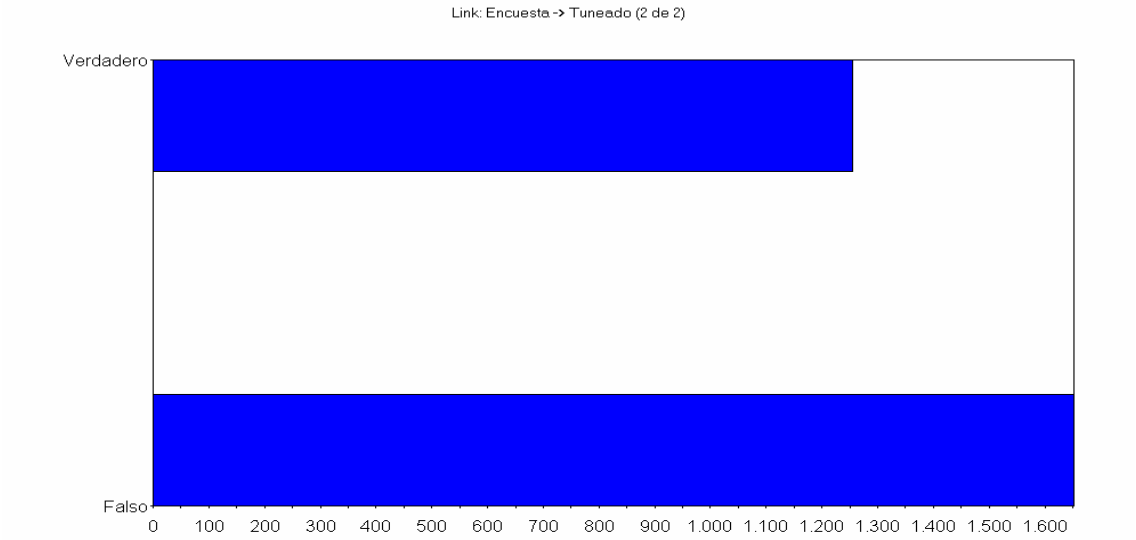


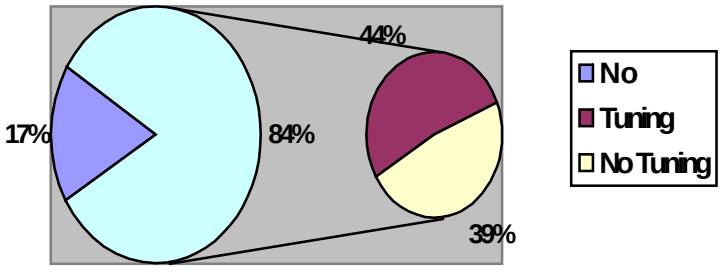
El que la PlayStation resulti la videoconsola més posseïda era predictible, però es interessant que veure que a més de la més posseïda ho és de manera molt contundent..

Coche

La gran majoria dels enquestats tenen cotxe, lo que resulta interessant, ja que haurem de pressuposar que els que s'interessen pel tuning ja tenen un cotxe. També apuntem per a mirar després quants han començat a fer tuning al cotxe i quants dels que encara no tenen edat de tenir cotxe ja el tenen (algo que sembla estrany però que es comú entre els joves que treballen i ja volen tenir un cotxe que començar a cuidar; els menors de 16 podem considerar que es fan com a seu el cotxe del pare).



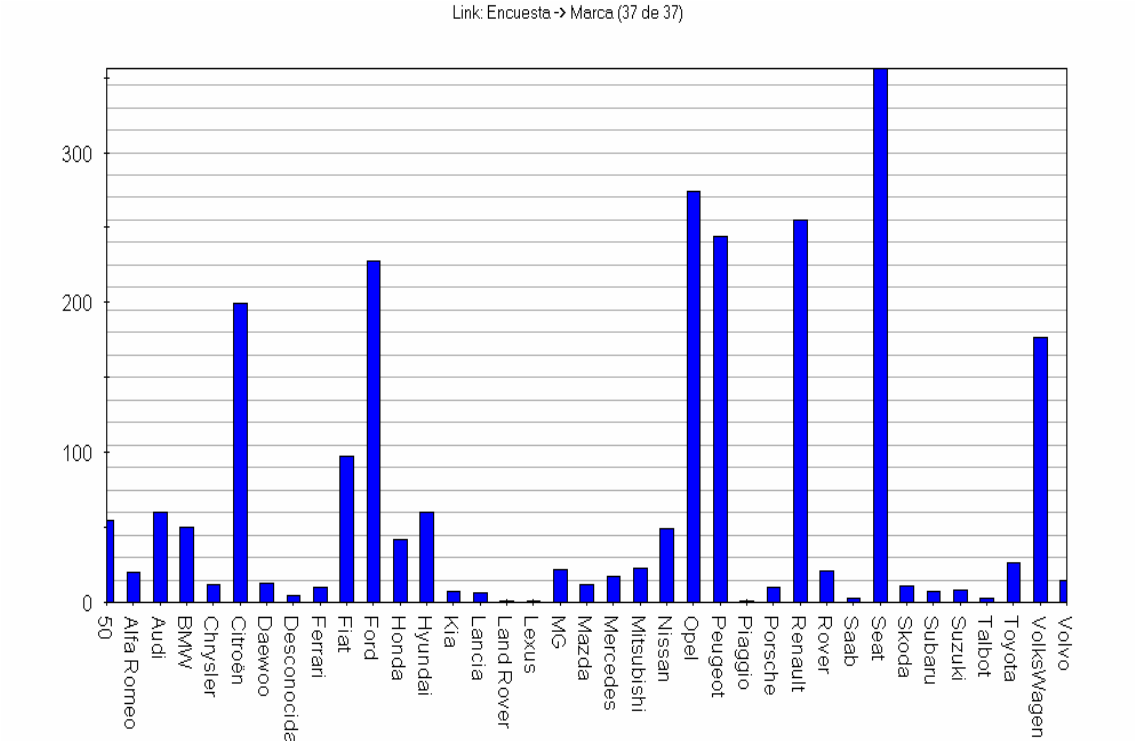


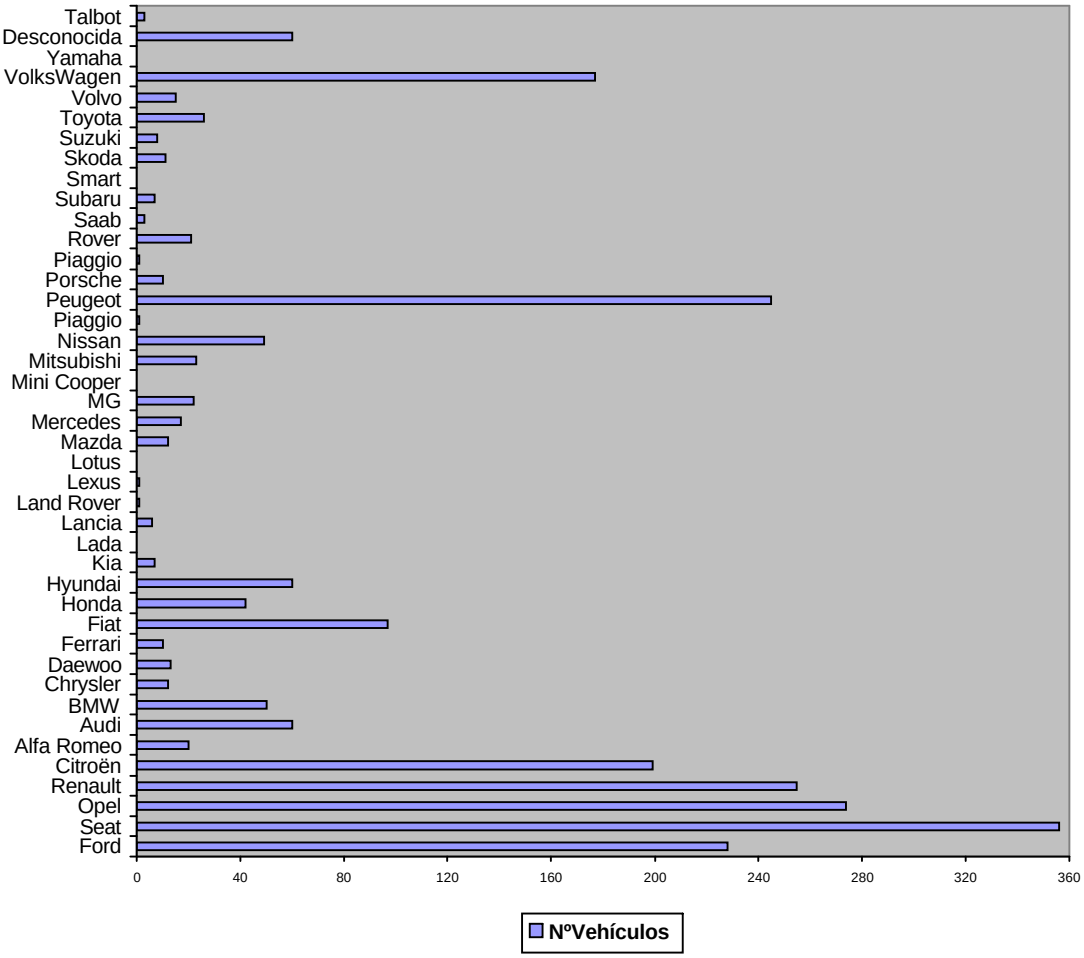


Marcas

Encara que a grans trets reflecteix la realitat dels cotxes venguts a la població global, hi ha que tenir en compte la presencia de bastants vehicles de gamma alta, que són reflex del

poder adquisitiu de la gent que s'interessa per la transformació dels vehicles (sobretot del motor en aquests casos).

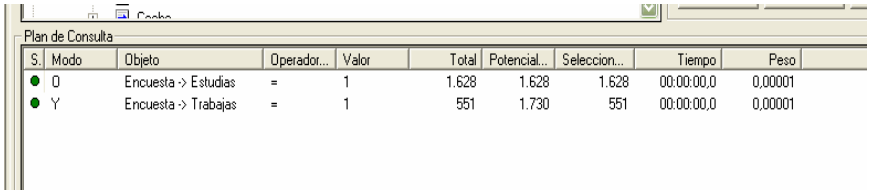




CONSULTES:

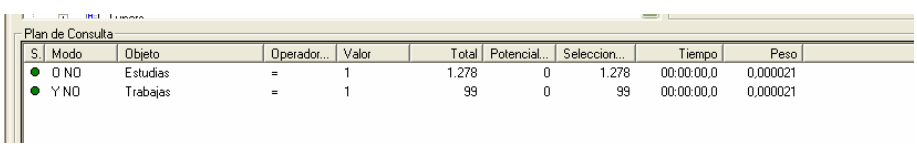
Usant l'opció de consultes de Synera Explorer en faig algunes d'inicials d'on descobreixo que:

- Els que estudien i treballen alhora son una part relativament important: 551 persones, un 18,96%:

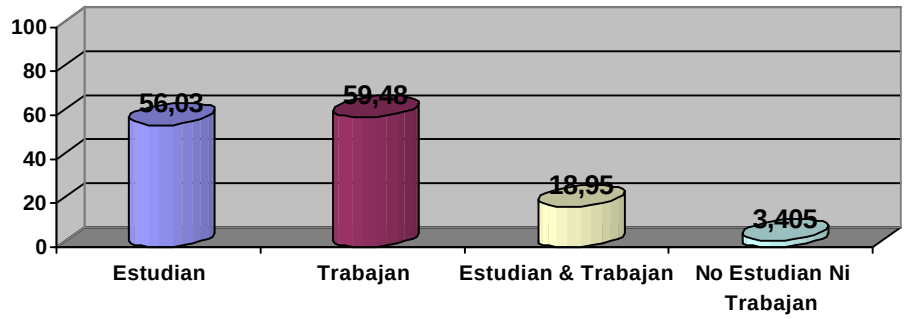

Plan de Consulta

S.	Modo	Objeto	Operador...	Valor	Total	Potencial...	Seleccion...	Tiempo	Peso
●	O	Encuesta -> Estudias	=	1	1.628	1.628	1.628	00:00:00,0	0,00001
●	Y	Encuesta -> Trabajas	=	1	551	1.730	551	00:00:00,0	0,00001

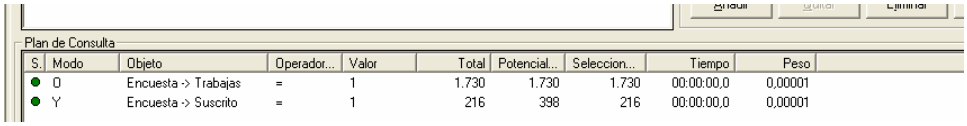
- Els que no estudien ni treballen son una minoria, 99, un 3,40%:


Plan de Consulta

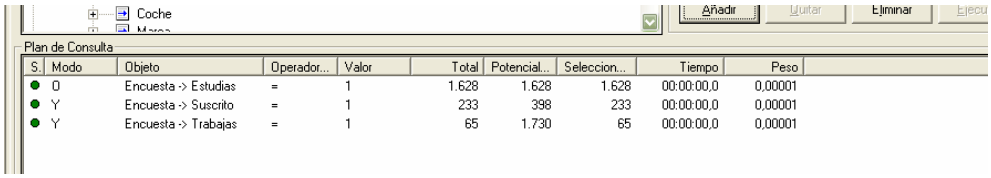
S.	Modo	Objeto	Operador...	Valor	Total	Potencial...	Seleccion...	Tiempo	Peso
●	O NO	Estudias	=	1	1.278	0	1.278	00:00:00,0	0,000021
●	Y NO	Trabajas	=	1	99	0	99	00:00:00,0	0,000021



- Significatiu és que dels 398 que estan inscrits, 216 treballen, 233 estudien i 65 estudien y treballen

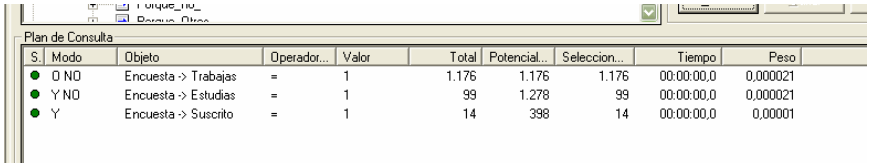

Plan de Consulta

S.	Modo	Objeto	Operador...	Valor	Total	Potencial...	Seleccion...	Tiempo	Peso
	0	Encuesta -> Trabajas	=	1	1.730	1.730	1.730	00:00:00,0	0,00001
	Y	Encuesta -> Suscrito	=	1	216	398	216	00:00:00,0	0,00001


Plan de Consulta

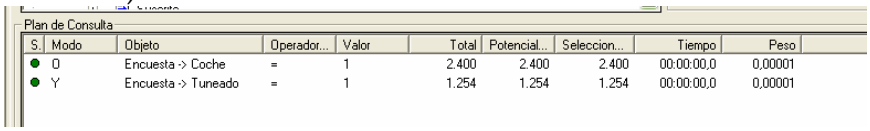
S.	Modo	Objeto	Operador...	Valor	Total	Potencial...	Seleccion...	Tiempo	Peso
	0	Encuesta -> Estudias	=	1	1.628	1.628	1.628	00:00:00,0	0,00001
	Y	Encuesta -> Suscrito	=	1	233	398	233	00:00:00,0	0,00001
	Y	Encuesta -> Trabajas	=	1	65	1.730	65	00:00:00,0	0,00001

En el cas contrari (que no estudiïn ni treballin però estiguin inscrits el cas no arriba al 5% dels inscrits:


Plan de Consulta

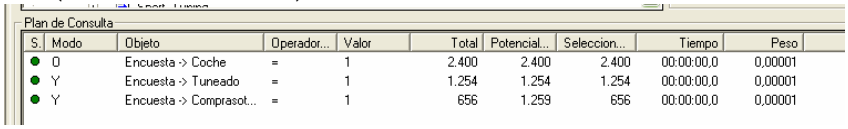
S.	Modo	Objeto	Operador...	Valor	Total	Potencial...	Seleccion...	Tiempo	Peso
	0 NO	Encuesta -> Trabajas	=	1	1.176	1.176	1.176	00:00:00,0	0,000021
	Y NO	Encuesta -> Estudias	=	1	99	1.278	99	00:00:00,0	0,000021
	Y	Encuesta -> Suscrito	=	1	14	398	14	00:00:00,0	0,00001

- Dels 2400 que tenen cotxe 1254 el tenen tunejat (per aquest cas no calia fer consulta perquè lògicament tots els que contesten afirmativament a l'atribut tunejat tenen cotxe).


Plan de Consulta

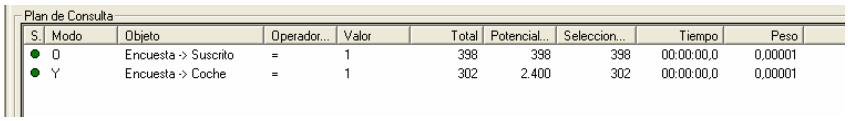
S.	Modo	Objeto	Operador...	Valor	Total	Potencial...	Seleccion...	Tiempo	Peso
	0	Encuesta -> Coche	=	1	2.400	2.400	2.400	00:00:00,0	0,00001
	Y	Encuesta -> Tuneado	=	1	1.254	1.254	1.254	00:00:00,0	0,00001

- Dels que tenen el cotxe tunejat un poc més de la meitat també compren altres revistes (dada interessant).


Plan de Consulta

S.	Modo	Objeto	Operador...	Valor	Total	Potencial...	Seleccion...	Tiempo	Peso
	0	Encuesta -> Coche	=	1	2.400	2.400	2.400	00:00:00,0	0,00001
	Y	Encuesta -> Tuneado	=	1	1.254	1.254	1.254	00:00:00,0	0,00001
	Y	Encuesta -> Comprasot...	=	1	656	1.259	656	00:00:00,0	0,00001

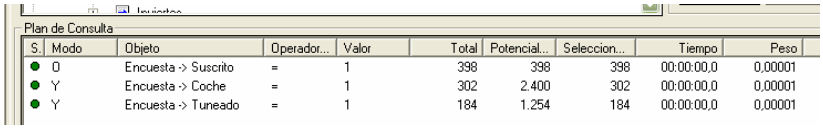
- Dels inscrits un tenen cotxe un 75% (dada interessant).



Plan de Consulta

S.	Modo	Objeto	Operador...	Valor	Total	Potencial...	Seleccion...	Tiempo	Peso
●	0	Encuesta -> Suscrito	=	1	398	398	398	00:00:00,0	0,00001
●	Y	Encuesta -> Coche	=	1	302	2.400	302	00:00:00,0	0,00001

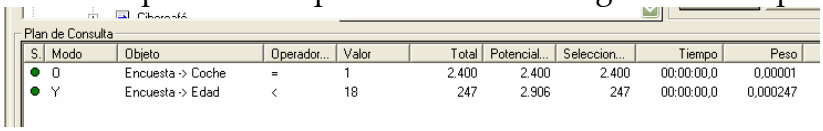
I un poc més de la meitat el tenen a més tunejat % tenen el cotxe tunejat (llavors la dada de si el tenen tunejat de moment no ens dona rellevant informació ja que es la mateixa proporció que els que tenen cotxe).



Plan de Consulta

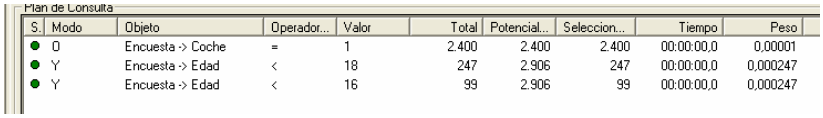
S.	Modo	Objeto	Operador...	Valor	Total	Potencial...	Seleccion...	Tiempo	Peso
●	0	Encuesta -> Suscrito	=	1	398	398	398	00:00:00,0	0,00001
●	Y	Encuesta -> Coche	=	1	302	2.400	302	00:00:00,0	0,00001
●	Y	Encuesta -> Tuneado	=	1	184	1.254	184	00:00:00,0	0,00001

- Dels 2400 que tenen cotxe, 247 tenen menys de 18 anys, y 99 en tenen menys de 16 (per tant aquests 99 no poden haver aconseguit el cotxe per ells mateixos.



Plan de Consulta

S.	Modo	Objeto	Operador...	Valor	Total	Potencial...	Seleccion...	Tiempo	Peso
●	0	Encuesta -> Coche	=	1	2.400	2.400	2.400	00:00:00,0	0,00001
●	Y	Encuesta -> Edad	<	18	247	2.906	247	00:00:00,0	0,000247



Plan de Consulta

S.	Modo	Objeto	Operador...	Valor	Total	Potencial...	Seleccion...	Tiempo	Peso
●	0	Encuesta -> Coche	=	1	2.400	2.400	2.400	00:00:00,0	0,00001
●	Y	Encuesta -> Edad	<	18	247	2.906	247	00:00:00,0	0,000247
●	Y	Encuesta -> Edad	<	16	99	2.906	99	00:00:00,0	0,000247

13.7. Synera Discovery

13.7.1. Introducció a Synera Discovery

El Discovery es l'aplicació que permet executar processos de Data Mining, en concret els mètodes:

- Segmentació segmentació (cluster): agrupació d'elements d'una població per la seva similitud en les seves característiques. La segmentació proporciona:
    - Objectivitat: els segments que crea no estan condicionats a idees prèvies i per tant poden ser segments tipològicament diferents als que hauriem creat manualment.
    - Completitud: al afegir posteriorment elements, aquests s'inclouran en algun dels segments creats, mentre que per divisió manual podrien quedar exclosos d'aquests segments.
- Anàlisis associatiu: el que es busca es trobar fets que impliquen la presencia d'altres fets en el mateix conjunt de valors de dades. Synera usa el MBA (Market Basket Analisys): SI A=x llavors B=y. Els conceptes que s'usen són:

Regles: Defineixen relacions d'implicació entre fets. La importància la donen el suport (nombre d'instàncies en les que el fet es vertader dividit pel nombre total=importància de la regla amb respecte al nombre total d'instàncies analitzades) i la confiança (probabilitat de que un fet passi com resultat de un altre que ha passat=grau de certesa de la regla donat un antecedent).
Nivell: és la propietat de les regles que diu el nombre de fets que intervenen en cada una de elles.

13.7.2. Tasques prèvies (també usant Explorer)

La categorització necessària per a convertir els valors de un link pot realitzar-se de manera manual (com veurem en aquest apartat) o de manera automàtica mitjançant l'algoritme d'extrems o el de cluster.
Mentre que la tècnica de cluster es realitza com hem vist anteriorment la d'extrems: s'agafen valors des del mínim cap al màxim fins que el nombre de valors sigui igual al suport demanat. Després s'agafen valors del màxim cap al mínim fins que el nombre de valors sigui igual al suport especificat.

Els atributs on encara existeixin valors nuls (d'introducció lliure de text) els tractaré de manera diferent segons si vull tractar la manca de valors com una dada més (llavors els agruparé com un valor més) o si no vull tindre'ls en compte en l'anàlisi (llavors els obviaré).

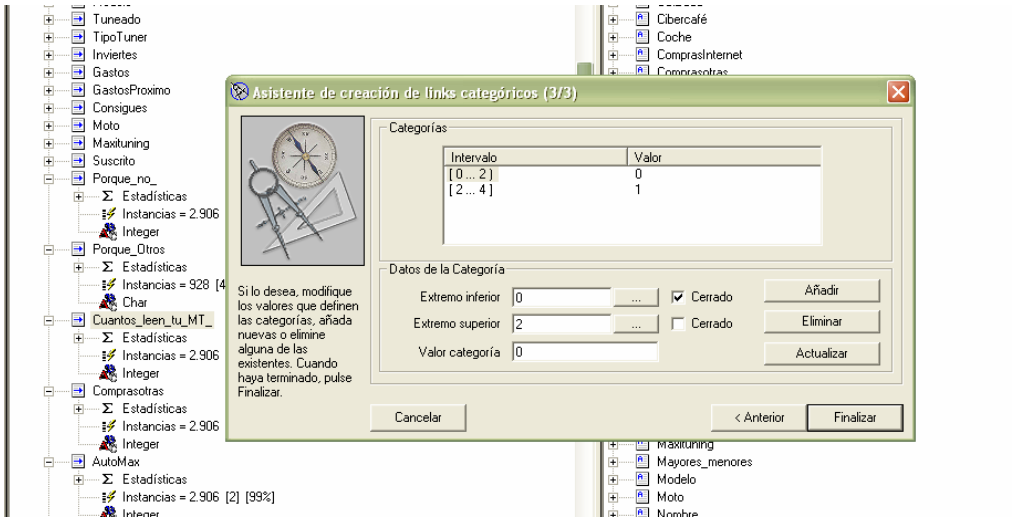
Categorització:

Edad: com que vull tractar els menor de edat per un costat i els majors per l'altre (per a diferenciar dos línies, quins tenen cotxe sense haver complit els 18 i com es diferencien els dos grups) faig una de edat 1 a 17 (obert) i de 18 (tancat) a 95 (obert)

Intentaré més tard pel Discovery que hem faci automàticament dos grups, però no ho aconsegueixo (passo l'atribut a enter i a char i tampoc hi ha manera).

Però després en faig també una altra, que potser també em serà útil: de 16 a 19 y de 19 a 41 (llevant una desviació estàndard de cada extrem)

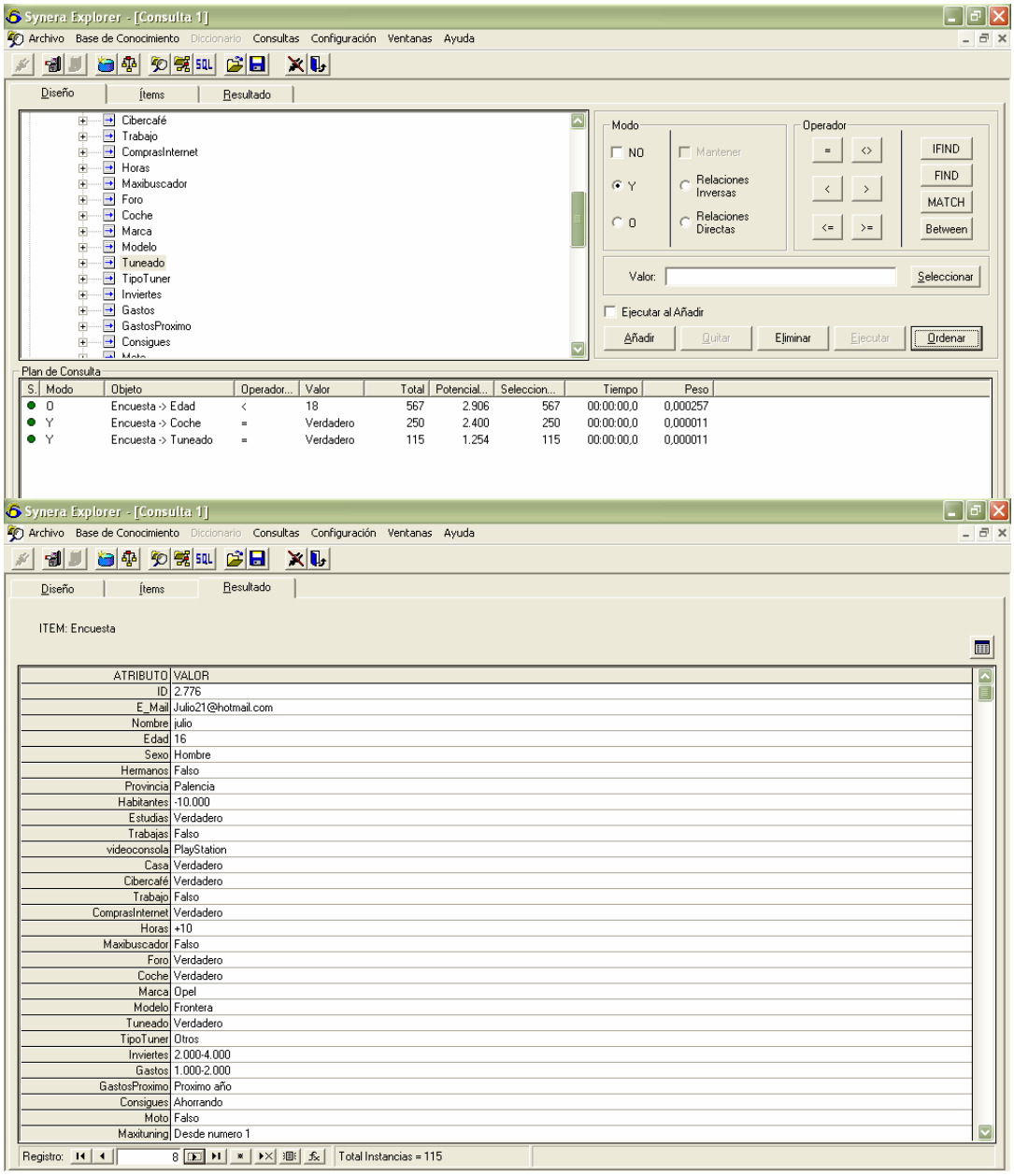
Cuantos leen tu MT: divideixo les opcions en dos grups.



Hores que es passen a internet: Horas:
De 1 a 4 (que seria de 2 a 8 hores) ho poso tot en la categoria de valor 0 y la 5 (més de 10) a al categoria 1.

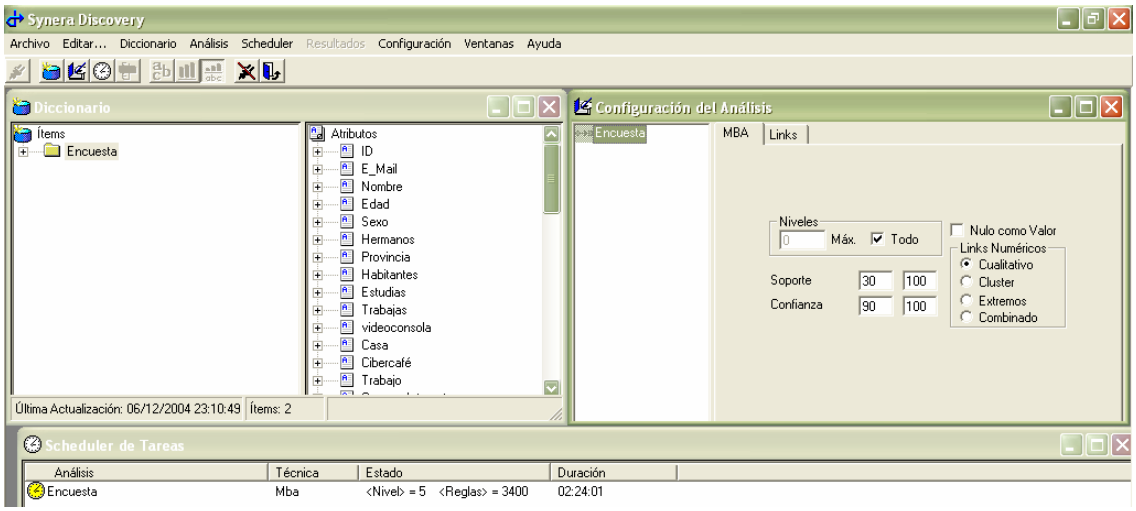
Quan s'implanti Synera a l'empresa la gestió d'Usuaris que podran modificar y carregar bases s'haurà de completar. De moment usarem les opcions de Synera que ens ajuden a treure coneixement de les dades. El coneixement (encara que en un principi intento el contrari) es treu incrementalment. A més es poden fer les tasques més bàsiques de tractament de bases de dades (sentències SQL).
Quan s'usin diversos ítems els recursos y la manera de treballar s'ampliarà moltíssim, per lo qual la feina de ensenyament del paquet no acabaria en la explicació al responsable dels passos que s'han efectuat aquí.
Y el que interessa en aquesta part es clar, trobar els patrons de comportament mitjançant tasques de Mineria de Dades.

Un cop aconseguides carregades les dades de la manera descrita al principi, tot amb un sol ítem i incloent els valors en els ítems "enters" que es relacionaven amb una altra taula torno a mirar amb l'Explorer els atributs, i encara que deixo per al final la búsqueda de les sentencies (el resultat de consultes) intensament per a refirmar el MBA, realitzo un parell de predictibles:



De moment, com ja he dit, no categoritzo cap atribut més enllà de la feina feta en el treball previ a l'anàlisi.

Després de fer un intent "salomònic" de començar trobant les associacions entre tots els atributs decideixo desistir i començar trobant associacions que ja m'han donat bons resultats en les consultes de l'Explorer.



Però per més que intento ara buscar relacions entre les dades que vagin més enllà de les visibles (les que ja són obvies per ser els valors majoritaris) els resultats són els obvis, i encara que canviï les mesures de confiança i suport per a ajustar els resultats, tota la informació que extrec és la que ja he trobat al principi de la feina, quan feia un anàlisi previ coneixent les dades.

Finalment torno enrera i treballo amb sentències SQL, però encara que intento trobar algun indici de relació diferent a les obvies per a poder treballar-la després no hi ha manera. Igualment hi torno amb les consultes, però al tenir només un ítem quan intento complexar-ho fent operacions entre elles hem dona el mateix resultat. Les observacions que a primera vista vaig fer a partir dels gràfics inicials de les dades: no trobo cap informació rellevant que hem sigui útil i interessant.

Per una altra línia netejo (amb menys eficàcia) l'altra base de dades i la deixo en 4.340 entrades. Realitzo les tasques prèvies i els resultats són si fa no fa els mateixos. Finalment desisteixo.

Per tant (i encara que les relacions trobades són molt simples) acabo fent les següents afirmacions:

L'edat dels potencials lectors és, per contra de l'esperat, està concentrada dels 17 als 24 anys, sent els 20 la mitjana. Son majoritàriament homes i la distribució per comunitats pot aproximar-se a la real exceptuant tres grans concentracions: Madrid, Barcelona i València, que destaquen. La combinació de estudiant-treballador és més freqüent que en la població total –però no destacable-, com també la possessió de cotxes d'alta gamma per sobre dels de gamma mitja. Es gasten de 600 a 1000 € (encara que aquesta afirmació no es contundent, com tampoc les despeses en hobbys).
Dels que tenen cotxe hi apliquen (lògicament) tuning moltíssims més que la població total (un 44% contra el 1% que realment hi ha tunejats a Espanya). El tipus més escollit el fino (però no manté relació ni amb l'edat, ni amb els altres atributs)

No estan subscrits sobretot per preferir comprar la revista al quiosc a esperar a rebre-la encara que també pesa el no disposar dels diners de cop (els menors contesten majoritàriament perquè els pares no els deixen), i part important (75%) tenen cotxe, i aquí els tunejats segueixen la mateixa proporció que la totalitat de inscrits-no inscrits.
La revista més comprada és GTI, i el maketuning no pren rellevant interès. Per contra de la F1, que si ho fa. No hi ha relació entre les revistes que es compren o no, encara que molts dels compradors de GTI també ho son de AutoMax.

14. RESENYES

http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html
http://energiaycomputacion.univalle.edu.co/edicion17/revista17_1a.BAK
http://www.gestiopolis.com/canales3/ger/gesdds.htm
http://dmi.uib.es/~bbuades/datawarehouse/sld003.htm
http://www.monografias.com/trabajos/datamining/datamining.shtml
curs KDD: http://dns1.mor.itesm.mx/~emorales/cursos/KDD
https://www.idg.es/computerworld/noticia.asp?id=10125
http://www.profinmexico.com/boletines/*
http://answermath.com/*
Comunicaciones en Socioeconomía, Estadística e Informática. 2003. Vol.7 Num.2.
Proyecto CONACYT 125939-B
"Redes Neuronales y Sistemas Borrosos", Bonifacio Martín del Brío y Alfredo Sanz, ed.RA-MA 1997.
Quintas, Paul; Lefrere, Paul; Jones, Geoff, "Knowledge Management: a Strategic Agenda", Long Range Planning, Vol. 30, No. 3, pp. 385 a 391, 1997, Elsevier Science Ltd.

15. GLOSSARI

Data Warehouse: repositori central amb la informació més valuosa de l'empresa, on s'emmagatzemen les dades estratègiques, tècniques i operatives. Les dades emmagatzemades aquí han passat un procés de qualitat que assegura la seva consistència. A més el repositori esta construït per a que l'accés sigui lo mes ràpid possible.
La seva construcció es fa per etapes que normalment es corresponen a les principals àrees operatives de l'empresa. Estes àrees reben el nom de **Data Marts**.

Data-Mars: Repositori parcial de dades de l'empresa, on s'emmagatzemen les dades tàctics i operatius, per a obtenir informació tàctica.

Data Mining: Procés que ajuda a descobrir els patrons i relacions que poden passar desapercebuts en l'anàlisi del negoci i els clients. Ha d'estar orientar a resoldre un problema de negocis i no ha de necessitar el ser un especialista en la matèria per a poder usar-lo.

Gestió del coneixement: procés continu d'adquisició, distribució i anàlisi de la informació que es mou en l'entorn de l'organització per a fer més intel·ligent als seus treballadors (que cundeixin més), i ser més precisos en la presa de decisions, donar una resposta més ràpida a les necessitats del mercat i ser més competitiu en aquest entorn tan canviant.

Knoweldge Discovery in Databases: és el procés complet d'extracció d'informació, que s'encarrega a més de la preparació de les dades i la interpretació dels resultats obtinguts.

EIS (Executive Information System): Eines per a proveir d'informació estratègica als executius, mitjançant informes, comparatives y quadres de control multi-dimensionals.

DSS(Decission Suport System): Eines de decisió y anàlisis de dades no predefinides en les possibilitats d'un EIS, que entre altres característiques, han de ser senzilles, manejables i entendibles per a poder ser usades pel usuari que ha de decidir.

OLTP(On-Line Transaction Processing): Defineix el comportament habitual d'un entorn operacional de gestió:

- Altes/Baixes/Modificacions/Consultes
- Consultes ràpides y curtes
- Poc volumen d'informació
- Transaccions ràpides
- Gran nivell de concurrència

OLAP: On-Line Analytical Process: Defineix el comportament d'un sistema d'anàlisis de dades i elaboració d'informació:

- Només consulta.
- Consultes pesades i no predictibles
- Gran volum d'informació històrica
- Operacions lentes.

Es distingeix entre MOLAP I ROLAP (multidimensional i relacional)