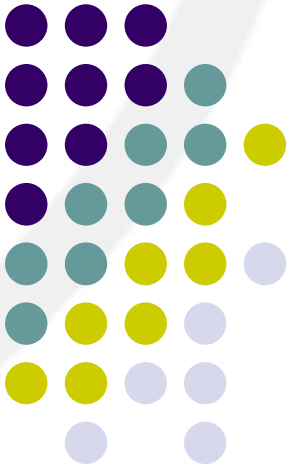
Progetto “COOL PATTERNS”

Analisi delle vendite nella grande distribuzione

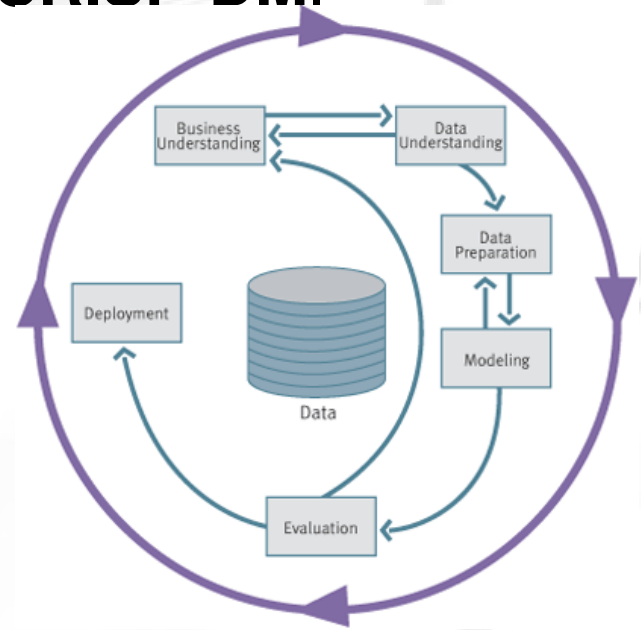
Analisi dei Dati ed
Estrazione di Conoscenza
2004/2005

Federico Colla



Introduzione

- Il progetto è stato condotto seguendo le linee guida della metodologia **CRISP-DM**.

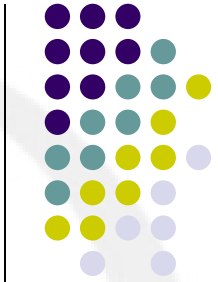


- Non tutte le fasi/task del **CRISP-DM** sono state trattate per via della natura didattica del progetto.

Strumenti di analisi

- Software:
  - *SPSS Clementine 8.1* (regole associative, alberi di decisione, esplorazione e preparazione dati)
  - *PrefixSpan_O*: una implementazione freeware dell'algoritmo Prefix Span (pattern sequenziali)
  - *Microsoft Visual Studio .NET 2003*
  - 5 programmi sviluppati appositamente per il progetto
  - *EDXOR* e alcuni script di supporto
- Hardware:
  - Notebook Intel Pentium IV 2.66Ghz 512MB RAM

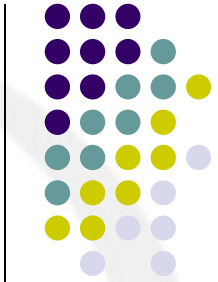
Business understanding



- Analisi delle vendite nella grande distribuzione.
- I dati disponibili sono relativi alle vendite di un ipermercato, effettuate nel trimestre Gennaio–Marzo del 2005.

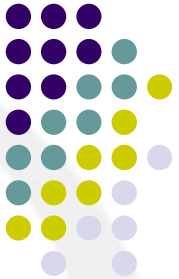
- Due obiettivi:
  - Trovare associazioni *interessanti* tra prodotti venduti insieme e sequenze tipiche di prodotti acquistati, a diversi livelli di astrazione.
  - Estrazione di un profilo dei clienti che supportano le regole/sequenze di acquisto trovate nella fase precedente.

Business Understanding:
Business objectives



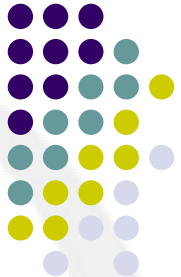
- Vogliamo poter rispondere alle seguenti domande:
  - Quali sono i prodotti il cui acquisto influenza l'acquisto di altri prodotti?
  - Quali sono i prodotti che vengono spesso comprati in sequenza nel tempo?
  - Qual'è l'*identikit* del cliente il cui comportamento di acquisto soddisfa un certo pattern?
- **Success criteria**
  - Si chiede che le associazioni/sequenze di prodotti trovate siano *interessanti*
  - Cosa vuol dire *interessanti*?

Business Understanding:
Data mining goals



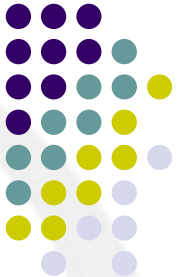
- Primo obiettivo – Analisi delle vendite
  - Regole associative a singola dimensione e multilivello.
  - Pattern sequenziali di prodotti venduti nel tempo, a diversi livelli di astrazione.
- Secondo obiettivo – Estrazione profilo clienti
  - Per ogni regola associativa/pattern sequenziale interessante
    - Crea un albero di decisione che classifica i clienti rispetto ad un attributo target il quale è positivo se il cliente soddisfa la regola associativa/pattern sequenziale, e negativo altrimenti.
    - L'analisi dell'albero risultante permette l'estrazione del profilo di interesse.

Data Understanding:
Data collection



- L'insieme dei dati iniziali è rappresentato da una serie di file di tipo e formato diverso:
  - `aggregato_scontrini_iper_liv.lst` scontrini aggregati (171MB)
  - 78 file `iva_scontrini_AAAAMMGG.lst` che rappresentano gli scontrini di ogni giornata lavorativa del trimestre (525MB)
  - `anag_articoli.lst` dati anagrafici relativi agli articoli
  - `carte_liv.lst` dati anagrafici relativi ai soci
  - `Classificazione Marketing.xls` classificazione gerarchica dei prodotti
  - `tracciati.txt` descrizione del formato dei vari file

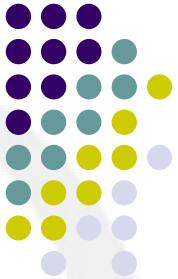
Data Understanding:
Data description – Scontrini dettagliati



- File di testo con record a lunghezza fissa di 63 caratteri, comprendente i seguenti 8 campi:
  - *data_scontrino*: char(10) con formato ‘gg/mm/aaaa’
  - *cassa*: number(3) numero cassa
  - *scontrino*: number(4) numero scontrino
  - *cod_categoria*: char(3) codice categoria
  - *cod_art* : number(6) codice articolo
  - *venduto_valore*: number(13) valore spesa
  - *qta_pezzi* : number(10) quantit`a pezzi
  - *qta_peso*: number(13) quantit`a peso

- Esiste un record per ogni prodotto acquistato, e la tripla *(data_scontrino, cassa, scontrino)* identifica univocamente il carrello di appartenenza.
- Il caricamento dei dati relativi agli scontrini dettagliati avviene attraverso un nodo di input per file a lunghezza fissa di *Clementine*.

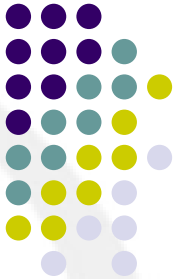
Data Understanding:
Data description – Scontrini aggregati



- File di testo con record a lunghezza fissa di 69 caratteri, comprendente i seguenti 8 campi
  - *data_scontrino*: char(10) con formato ‘gg/mm/aaaa’
  - *cassa*: number(3) numero cassa
  - *scontrino*: number(4) numero scontrino
  - *nro_carta*: number(8) numero carta socio
  - *cod_aggregato*: number(4) giustapposizione settore/reparto
  - *venduto*: number(13) valore spesa
  - *venduto_promo*: number(13) valore spesa prodotti in promozione
  - *venduto_pam*: number(13) valore spesa prodotti a marchio COOP

- La tripla *(data_scontrino, cassa, scontrino)* identifica univocamente il carrello di appartenenza.
- Permettono di legare gli scontrini con il numero di carta socio e quindi risalire alle informazioni riguardanti il cliente.
- Il caricamento dei dati relativi agli scontrini dettagliati avviene attraverso un nodo di input per file a lunghezza fissa di Clementine.

Data Understanding:
Data Description – Anagrafica prodotti

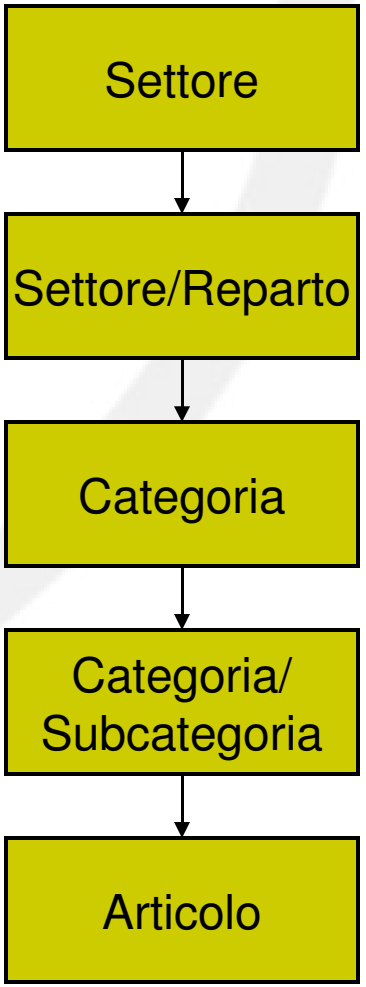


- File di testo con record a lunghezza fissa di 53 caratteri, comprendente i seguenti 6 campi
  - *cod_art* : number(7) codice articolo
  - *art_descr*: char(30) descrizione articolo
  - *cod_presmkt*: char(1) vendibilità commerciale
  - *cod_clmkt*: char(11) codice classificazione marketing
  - *cod_clgest*: char(3) codice categoria marketing
  - *cod_stat*: char(1) stato articolo

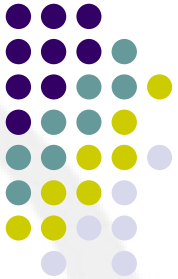
- Il codice di classificazione marketing (*cod_clmkt*) permette di ricavare la descrizione dell'articolo a diversi livelli di astrazione.
- Il codice "SSRRCCCZZXX" ha la seguente interpretazione:
  - **SS** è il codice del settore
  - **RR** è il codice del reparto
  - **CCC** è il codice della categoria
  - **ZZ** è il codice della subcategoria
  - **XX** non è utilizzato

Data Understanding:
Data description – Gerarchia prodotti

- La descrizione della gerarchia degli articoli è specificata nel file Excel `Classificazione Marketing.xls`.
- Si estraggono 4 tabelle che descrivono ciascuna un livello della gerarchia (chiave, descrizione)
  - **Settori**
    - 9 record, 2 campi (chiave: *cod_settore)*
  - **Reparti**
    - 54 record, 3 campi (chiave: *cod_settore + cod_reparto)*
  - **Categorie**
    - 402 record, 4 campi (chiave: *cod_categ*)
  - **Subcategorie**
    - 1 516 record, 5 campi (chiave: *cod_categ* + *cod_subcateg*)



Data Understanding:
Data description – Anagrafica soci

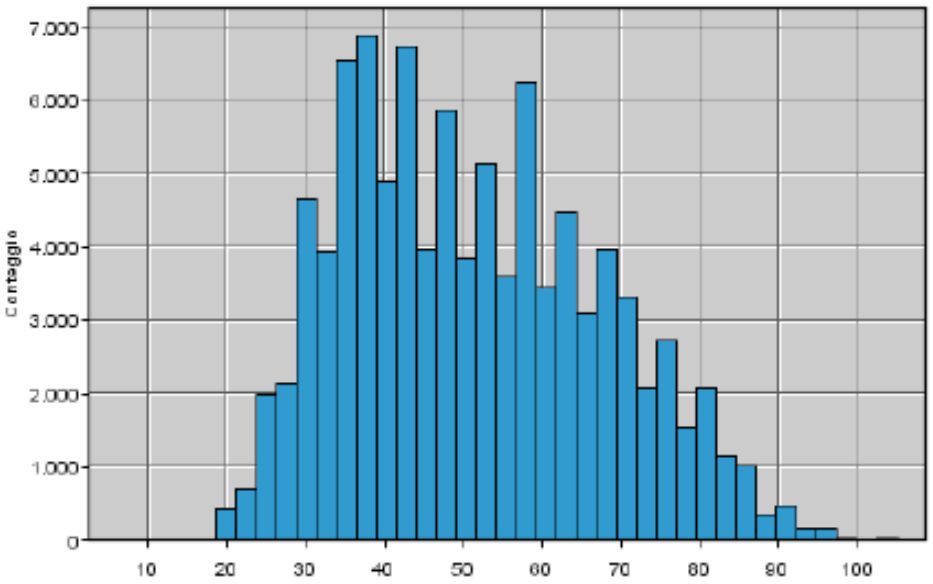


- File di testo con record a lunghezza fissa di 189 caratteri comprendente i seguenti 10 campi:
  - *nro_carta*: number(10) numero carta socio
  - *nro_socio*: number(5) numero socio
  - *data_nasc*: char(10) con formato ‘gg/mm/aaaa’
  - *sesso*: char(1)
  - stato_civile: char(1)
  - *professione*: char(50)
  - *titolo_studio*: char(50)
  - *res_citta*: char(50) residenza
  - *res_cap*: char(5) CAP residenza
  - *socio_capofam*: number(5)

- Questi dati saranno usati per la costruzione dei classificatori per l’estrazione dei profili dei clienti.

Data exploration – Dati Clienti

- A partire dall'attributo *data_nasc* possiamo ricavare l'età del cliente



- L'attributo *age* ha minimo 6, massimo 105 e media 51.

Data exploration

Valore	Proporzione %	Conteggio
F	61,544	60 198
M	38,456	37 616

Valore	Proporzione %	Conteggio
C	31,811	31 116
D	0,347	340
I	0,926	906
P	0,563	551
S	65,152	63 728
V	1,199	1 173

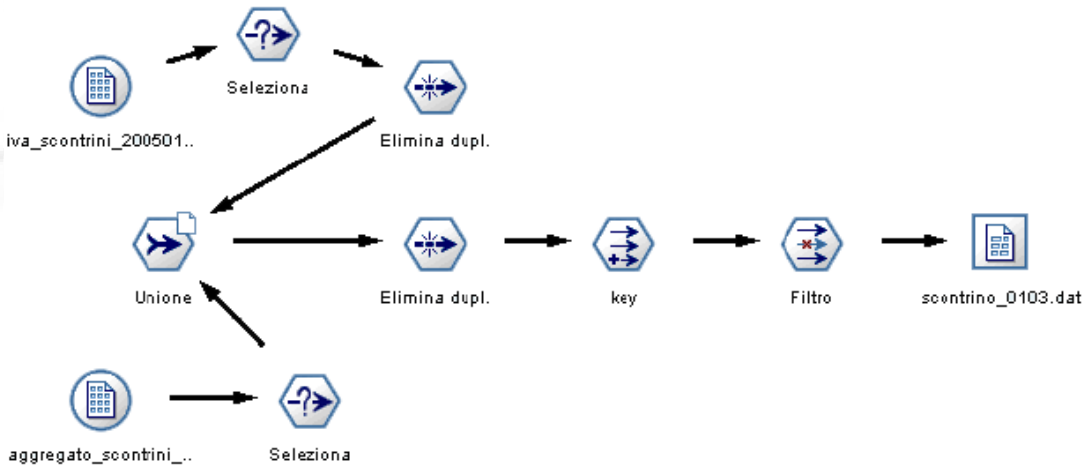
Data exploration

Valore	Proporzione %	Conteggio
AGRARIA	0,009	9
ALBERGHIERO	0,022	22
CLASSICO	0,106	104
ECONOMIA E COMM.	0,027	27
ELEMENTARE	2,961	2 897
GEOMETRA	0,259	254
GIURISPRUDENZA	0,038	38
INFORMATICA	0,0163	16
INGEGNERIA	0,077	76
LAUREA - ALTRO	3,151	3 083
LAUREA BREVE	0,001	1
MAGISTRALE	0,788	771
MATER. CLASSICHE	0,020	20
MATURITA' ALTRO	2,925	2 862
MEDIA INFERIORE	8,821	8 629
MEDIA SUPERIORE	6,438	6 298
NON INDICATA	72,1278	70 551
RAGIONERIA	1,487	1 455
SCIENTIFICO	0,359	352
SCIENZE ECONOM.	0,001	1
SCUOLA PROFESSIONALE	0,034	34
TECNICO AGRARIO	0,023	23
TECNICO INDUSTR.	0,276	270
TURISTICO	0,021	21

Valore	Proporzione %	*Conteggio*
AGRICOLTORE	0,279	273
ALTRE PROFESSIONI	0,093	91
ARTIGIANO	2,190	2 143
CASALINGA	26,667	26 085
CLERO	0,0255	25
COMMERCIANTE	0,0368	36
COOP CONSUMO	4 0,239	234
DIRIGENTE DI AZIENDA	0,188	184
DISOCCUPATO	3,530	3 453
ENTE PRIVATO	0,905	886
ENTE PUBBLICO	5,877	5 749
IMPIEGATO	18,191	17 794
INSEGNANTE	0,047	46
LAVORATORE AUTONOMO	3,453	3 378
LIBERO PROFESSIONISTA	3,282	3 211
MILITARE DI CARRIERA	1,374	1 344
NON INDICATA	1,120	1 096
OPERAIO	16,2087	15 854
PENSIONATO	9,162	8 962
POSSIDENTE	0,008	8
STUDENTE	7.117	6 962

Data preparation – Obj 1
Dataset construction – Regole associative

- Preprocessing scontrini dettagliati (78 run):

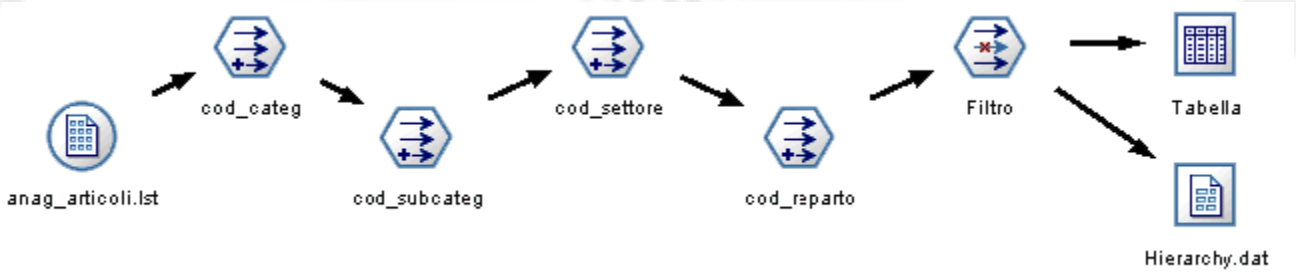


- Join con gli scontrini aggregati per associare lo scontrino al cliente (*nro_carta*)
- Vengono eliminati i record duplicati e quelli aventi il codice articolo ‘174292’ (la busta della spesa)
- *key* = MMGG + *cassa* + *scontrino*
  - MM = mese, GG = giorno

Data preparation – Obj 1
Dataset construction – Regole associative

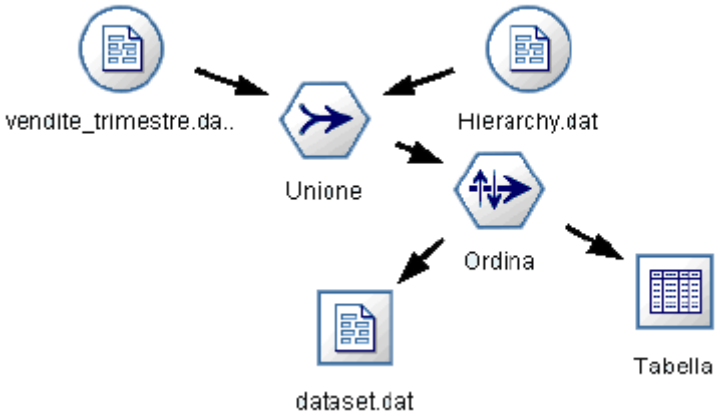
- I file risultanti sono stati concatenati (mantenendo l'ordine cronologico) in un unico file `vendite_trimestre.dat`

- La gerarchia dei prodotti viene creata a partire dal campo *cod_clmkt* contenuta nell'anagrafica articoli



Data preparation – Obj 1
Dataset construction – Regole associative

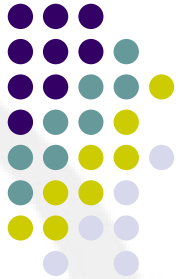
- Infine si crea di dataset finale (per le regole associative)



- Contiene 5 098 533 record e 7 campi
- Estratto:

key	nro_carta	cod_art	cod_categ	cod_subcateg	cod_reparto	cod_settore
01030011731	31403686	2561	009	01	01	01
01030011731	31403686	2545	009	01	01	01
01030011731	31403686	3393	009	03	01	01

Data preparation – Obj 1
Dataset construction – Pattern sequenz.



- Il software PrefixSpan_O richiede in input un file binario
- Ad esempio le sequenze

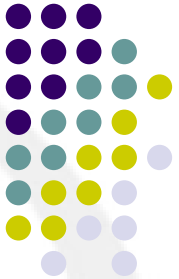
{ (0 1) (2) (1 3) }, { (1) (3) }

- devono essere trasformate in

0 1 –1 2 –1 1 3 –1 –2 1 –1 3 –1 –2

- Gli item sono numerati da 0 a |*items*|-1 (interi di 32 bit)
- "–1" delimita le transazioni dello stesso cliente
- "–2" delimita le sequenze di clienti diversi

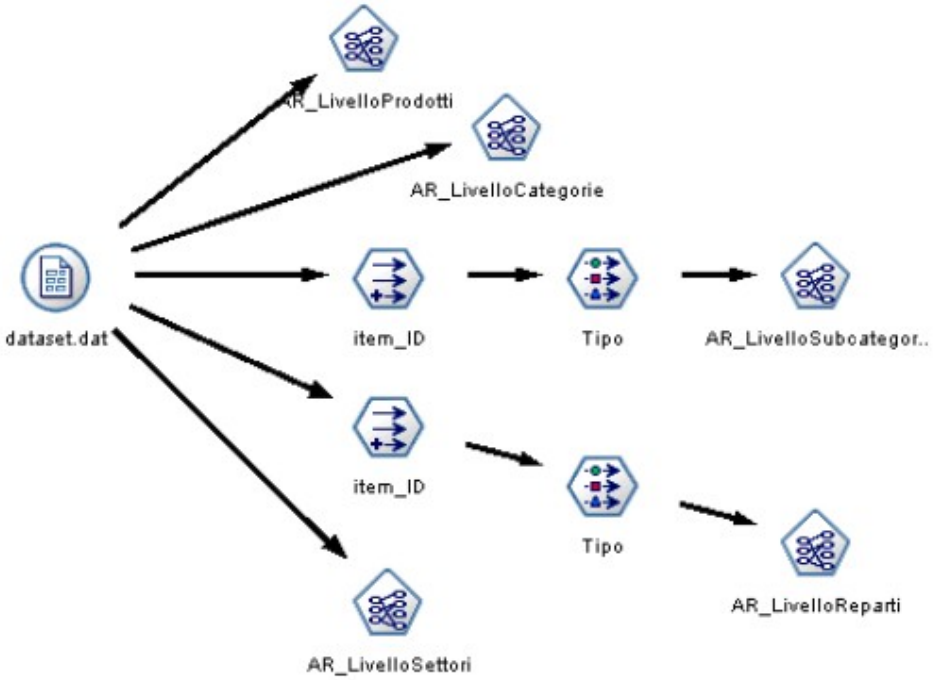
Data preparation – Obj 1
Dataset construction – Pattern sequenz.



- Per creare il file di input è stato scritto un programma, *DSBuilder*, che dati
  - la lista ordinata degli item che compaiono *effettivamente* nelle transazioni e
  - il file `dataset.dat`,
- restituisce in output un file binario, formattato come descritto in precedenza, che contiene le sequenze di acquisto di tutti i clienti, al livello di astrazione desiderato.
- Vengono creati 5 dataset, ciascuno ad un livello di astrazione diverso.

Modeling – Obj 1
Estrazione regole associative

- *Clementine* ha permesso di effettuare l'analisi usando i dati in formato *transazionale*.
- L'attributo *key* identifica ogni transazione.
- A seconda del livello di astrazione considerato, i codici di articolo, subcategoria, categoria, reparto e settore sono gli attributi di input/output.



Modeling – Obj 1
Estrazione regole associative

- La strategia utilizzata per l'estrazione delle regole è quella del *reduced support*.
- Ogni livello di astrazione ha la sua soglia di supporto minimo
  - più basso è il livello nella gerarchia, più piccola è la soglia di supporto minimo corrispondente.

Livello	Supporto minimo	Confidenza minima
Articoli	0,01%	80%
Subcategorie	0,2%	75%
Categorie	0,7%	75%
Reparti	4%	75%
Settori	8%	80%

Regole interessanti → Lift maggiore di 1

Evaluation – Obj 1
Regole associative interessanti

- L'insieme di regole ottenuto è stato esportato in un file di testo in cui esiste un record per ogni regola

Istanze	Supporto	Confidenza	Lift	Conseguente	Antecedente 1
53	0.01	92.5	4237.263	283917	283920

- Le regole ottenute non sono direttamente interpretabili.
- E' stato scritto il programma *PrettyPrinterApriori* che, data una regola "grezza", restituisce la corrispondente descrizione testuale.

[10 BICCH.CART.BIBO CIRC.200CC] → [PIATTI CART.BIBO CIRCUS D23X10]

Line: 70 Support: 0,01 Confidence: 92,5 Lift: 4237,263

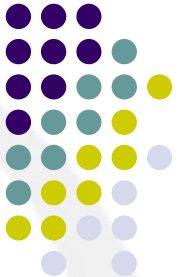
Evaluation – Obj 1
Regole associative – Articoli

- [10 BICCH.CART.BIBO CIRC.200CC] → [PIATTI CART.BIBO CIRCUS D23X10]
  - Support: 0,01 Confidence: 92,5 Lift: 4237,263
- [TELO 100X150 460 GR/MQ TU] [OSPITE 40X60 460 GR/MQ TU] → [ASCIUGAMANO 60X110 460 GR TU]
  - Support: 0,01 Confidence: 91,4 Lift: 965,993
- [BOCC.CANI POLLO/TACCH.KG1.23] [BOC/NI GATTO VITELLO SIM.KG415] → [BOCC.GATTI CONIGLIO SIMBA G415]
  - Support: 0,01 Confidence: 91,4 Lift: 390,042
- [PIATTO FRUTTA MAZIME B.CO CM21] [PIATTO F.DO MAXIME B.CO CM.17] → [PIATTO P.NO MAXIME B.CO CM.25]
  - Support: 0,01 Confidence: 90 Lift: 3052,386
- [LENZUOLO PIANO 150X280 RIGHE] [LENZUOLO ANGOLI 90X200 TU] → [FEDERA 50X80 STAMPA RIGHE]
  - Support: 0,01 Confidence: 87,8 Lift: 809,222

Evaluation – Obj 1
Regole associative – Articoli

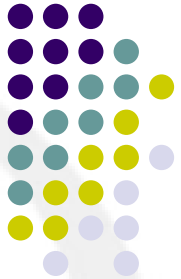
- [GOURM.GOLD DADINI GELLEE G85X8] [GOURMET PERLE FIL.C/MANZO G85] →[GOURMET PERLE FIL.CONIGLIO G85]
  - Support: 0,01 Confidence: 87,8 Lift: 492,757
- [CUCCHIAIONE ACCIAIO INOX] [PALA FRITTO ACCIAIO INOX] [FORCHETTONE ACCIAIO INOX] → [SCHIUMAROLA IN ACCIAIO INOX]
  - Support: 0,01 Confidence: 85,7 Lift: 1912,523
- [APER.CAMPARI MIXX PEACH ML275] [APERIT.CAMPARI MIXX LIME ML275] [APERITIVO CAMP.GRADI 6,5 ML275] → [CAMPARI MIXX ORANGE ML275]
  - Support: 0,01 Confidence: 83,3 Lift: 1314,55
- [GASSOSA S. BENEDETTO LT.1.5] [CEDRATA SAN BENEDETTO LT.1.5] [ARANCIATA S.BENEDETTO LT.1,5] → [SPUMA BIONDA LT1.5 S.BENEDETTO]
  - Support: 0,01 Confidence: 83 Lift: 172,76
- [BARAT.OVALE LT1,7 VTR COP.ACC.] [BARAT.OVALE LT0,84 VTR COP.ACC] → [BARAT.OVALE LT1,2 VTR COP.ACC.]
  - Support: 0,01 Confidence: 82,9 Lift: 1993,002
- [MOUSSE GAT.COOP MANZ/FEGAT.G85] [MOUSSE GAT.COOP PES/TROTAG85] → [MOUSSE GATTO COOP POL/TAC.G85]
  - Support: 0,1 Confidence: 81,7 Lift: 712,617

Evaluation – Obj 1
Regole associative – Subcategorie



- [BIBITE-ARANCIATE] [SNACK SALATI-PATATINE] [USA E GETTA TAVOLA-TOVAGLIE-TOVAGLIOLI] [USA E GETTA TAVOLA-STOVIGLIE PLASTICA BIANCA] → [BIBITE-COLE]
  - Support: 0,1 Confidence: 88,2 Lift: 11,084
- [USA E GETTA TAVOLA-STOV. CARTA COLORATA DECORATA] [USA E GETTA TAVOLA-ACCESSORI USA E GETTA] [USA E GETTA TAVOLA-STOVIGLIE PLASTICA BIANCA] → [USA E GETTA TAVOLA-TOVAGLIE-TOVAGLIOLI]
  - Support: 0,1 Confidence: 84,7 Lift: 12,767
- [USA E GETTA TAVOLA-STOV. CARTA COLORATA DECORATA] [USA E GETTA TAVOLA-STOV. PLAST. COLORATA DECORATA] → [USA E GETTA TAVOLA-TOVAGLIE-TOVAGLIOLI]
  - Support: 0,1 Confidence: 82,2 Lift: 12,391
- [SNACK SALATI-POP CORN/CEREALI] [SNACK SALATI-ESTRUSI] [BIBITE-ARANCIATE] → [BIBITE-COLE]
  - Support: 0,1 Confidence: 82,2 Lift: 10,34
- [CARAMELLE/PROD. BASE ZUCCH.-ALTRE CARAMELLE] [CARAMELLE/PROD. BASE ZUCCH.-CARAM.NORMALI] [CARAMELLE/PROD. BASE ZUCCH.-GOMME DA MASTICARE] → [PRODOTTI BASE CIOCCOLATO-SNACK]
  - Support: 0,1 Confidence: 81,2 Lift: 8,693

Evaluation – Obj 1
Regole associative – Categorie



- [UOVA] [OF PREPARATA] [VERDURA FRESCA] [LATTE] [FRUTTA FRESCA] → [ORTAGGI]
  - Support: 0,8 Confidence: 85,2 Lift: 1,893
- [CAFFE] [UOVA] [VERDURA FRESCA] [FRUTTA FRESCA] → [ORTAGGI]
  - Support: 0,7 Confidence: 84,3 Lift: 1,871
- [UOVA] [GRASSI] [VERDURA FRESCA] [AVICUNICOLO] → [ORTAGGI]
  - Support: 0,9 Confidence: 83,5 Lift: 1,854
- [OLIO DI OLIVA] [UOVA] [SUINO] → [BOVINO]
  - Support: 0,7 Confidence: 78,9 Lift: 1,757
- [ZUCCHERO] [IGIENE CARTA] [DETERGENTI SUPERFICI] → [DETERGENZA TESSUTI]
  - Support: 0,7 Confidence: 76,6 Lift: 2,247

Evaluation – Obj 1
Regole associative – Reparti

- FRESCHI-CARNI BIANCHE] [FRESCHI-SURGELATI] [FRESCHI-GASTRONOMIA] → [FRESCHI-CARNI ROSSE]
  - Support: 5,2 Confidence: 75,5 Lift: 1,217

- Al livello di Settore, non sono state trovate regole aventi Lift maggiore di 1.

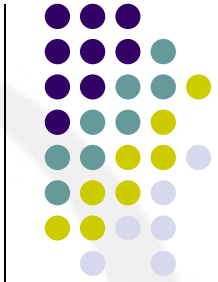
Modeling – Obj 1
Estrazione pattern sequenziali

- Per l'estrazione dei pattern sequenziali è stato utilizzato il software PrefixSpan_O
  - http://www-sal.cs.uiuc.edu/~hanj/software/prefixspan.htm.

- Strategia *reduced support*

Livello	Supporto minimo
Articoli	2%
Subcategorie	10%
Categorie	20%
Reparti	30%
Settori	40%

Evaluation – Obj 1
Pattern sequenziali interessanti



- Nel file di output esiste una linea per ogni pattern sequenziale.
- Un esempio di record ‘grezzo’ presente nel file di output è il seguente

(0 1) (3) : 0.321787

- Ogni gruppo di parentesi rappresenta un elemento della sequenza.
- Tra parentesi sono racchiusi gli identificativi degli item appartenenti alla stessa transazione.

- Il programma *PrettyPrinterPS* trasforma i pattern sequenziali ottenuti in una forma leggibile

Evaluation – Obj 1
Pattern sequenziali – Articoli

- { [BANANE SFUSE] } { [KIWI SFUSI] }
  - Support: 0,02218
- { [PATATE P.GIALLA COOP VTB KG2,5] } { [PATATE BOLOGNA CAL.4/7 R.KG2.5] }
  - Support: 0,023268
- { [PIZZAIOLA TRIS LOCAT.GR.125X3] } { [MOZZAR.S.LUCIA TRIS GR.125X3] }
  - Support: 0,010283
- { [FETTINE SCELTE VITELLONE] } { [PETTO POLLO COOP FETTE GL CF] }
  - Support: 0,010928
- { [IMPASTO PIZZA] } { [IMPASTO PIZZA] } { [IMPASTO PIZZA] }
  - Support: 0,007223

Evaluation – Obj 1
Pattern sequenziali

- **Livello Subcategorie**

- { [AVICUNICOLO-POLLO] } { [BOVINO-VITELLONE] }
  - Support: 0,148801
- { [FORMAGGI A SERV.ASSISTITO-SEMIDURI/DURI] } { [FORMAGGI LIBERO SERVIZIO-FORMAGGI FRESCHI] }
  - Support: 0,100328

- **Livello Categorie**

- { [BISCOTTI] } { [LATTE] }
  - Support: 0,222258
- { [FORMAGGI LIBERO SERVIZIO] } { [SALUMI LIBERO SERVIZIO] }
  - Support: 0,212621
- { [ORTAGGI] } { [VERDURA FRESCA] }
  - Support: 0,219369

Evaluation – Obj 1
Pattern sequenziali

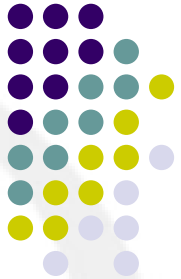
- **Livello Reparti**

- { [GENERI VARI-DROGHERIA ALIM. 1] } { [GENERI VARI-LIQUIDI] [FRESCHI-ORTOFRUTTA] }
  - Support: 0,381852

- **Livello Settori**

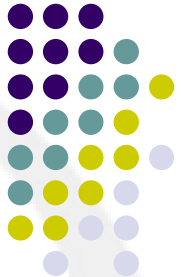
- { [GENERI VARI] [FRESCHI] } { [TESSILE] }
  - Support: 0,331645

Data preparation – Obj 2
Dataset construction



- Secondo obiettivo del progetto: estrazione di un profilo dei clienti che supportano le regole/sequenze trovate nella fase precedente.
- L'obiettivo di questo task è creare il dataset per l'estrazione dei profili dei clienti.
- Due programmi scritti appositamente per lo scopo
  - *CustomerIDRetrieverApriori*
  - *CustomerIDRetrieverPS*

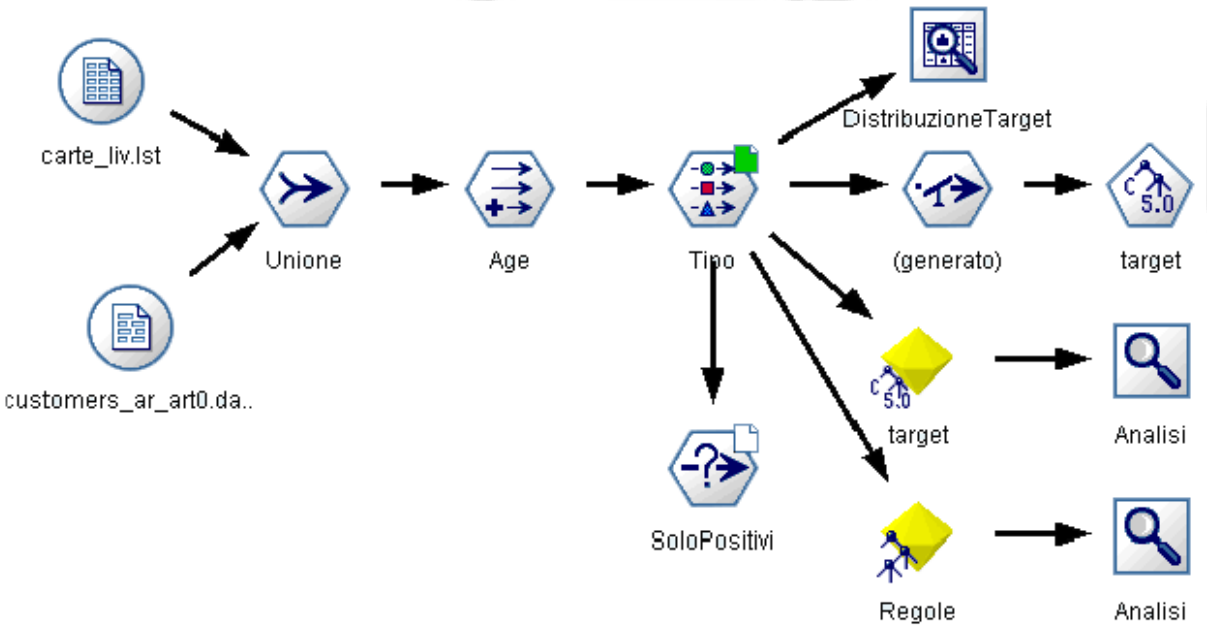
Data preparation – Obj 2
Dataset construction



- *CustumerIDRetrieverApriori* prende in input un file di testo contenente un insieme di regole associative 'grezze' e restituisce un file di testo contenente una tabella i cui record di due campi contengono:
  - Il numero della carta del cliente
  - Un flag binario che indica se il cliente supporta o meno la regola
- Il secondo programma, *CustumerIDRetrieverPS*, esegue la stessa operazione del precedente ma prende in input un file di testo contenente un insieme di pattern sequenziali 'grezzi'.
- Nei file di output esiste un record per ogni cliente che ha effettuato un acquisto nel trimestre considerato.
- I dataset finali sono creati in contemporanea alla fase di modeling.

Modeling – Obj 2

- Lo scopo di questa fase è quello di creare degli alberi di decisione per classificare la variabile target (il flag binario) definita in precedenza.
- L'algoritmo adottato per la costruzione dell'albero di decisione è il C5.0 (5-fold cross validation)



Modeling – Obj 2

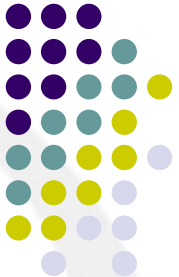
- Il dataset finale è una tabella che contiene i dati di *tutti e soli* i clienti che hanno effettuato acquisti nel trimestre.
- Ogni cliente ha associato un attributo binario (la variabile target).
- Attributi predittori:

Attributo	Tipo
sesso	flag
stato_civile	insieme discreto
professione	insieme discreto
titolo_studio	insieme discreto
age	intervallo

- A partire dall'albero di decisione ottenuto sono state generate le regole per la classificazione delle due classi.
- Per la creazione delle regole sono stati impostati livelli di confidenza minimi del 95%.

Evaluation – Obj 2

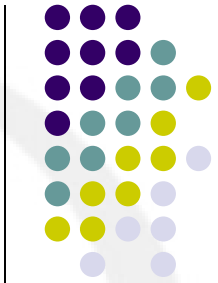
Regole associative – Articoli



- La regola
  - [GOURM.GOLD DADINI GELLEE G85X8] [GOURMET PERLE FIL.C/MANZO G85]→[GOURMET PERLE FIL.CONIGLIO G85]
  - Support: 0,01 Confidence: 87,8 Lift: 492,757
- è supportata da 41 clienti. Il classificatore ottenuto ha una accuratezza del 96,07% su tutti i dati, e del 100% sui dati classificati come positivi.
- Profilo cliente:
  - Casalinghe, non sposate, di età tra i 57 e i 63 anni, che hanno la terza media inferiore come titolo di studio.
  - Ragazze single ragioniere tra i 26 e i 29 anni che lavorano come impiegate
  - Uomini pensionati aventi età minore di 51 anni

Evaluation – Obj 2

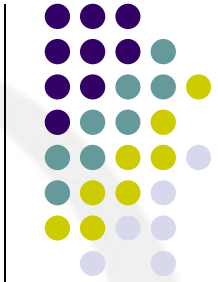
Regole associative – Articoli



- La regola
  - [APER.CAMPARI MIXX PEACH ML275] [APERIT.CAMPARI MIXX LIME ML275] [APERITIVO CAMP.GRADI 6,5 ML275] → [CAMPARI MIXX ORANGE ML275]
  - Support: 0,01 Confidence: 83,3 Lift: 1314,55
- è supportata da 27 clienti. Il classificatore ottenuto ha una accuratezza del 97,6% su tutti i dati, e del 100% sui dati classificati come positivi.
- Profilo cliente:
  - Ingegneri maschi aventi 26-27 anni
  - Ragazze dai 26 ai 30 anni che hanno un lavoro autonomo e sono diplomate
  - Impiegati single dai 26 ai 53 anni

Evaluation – Obj 2

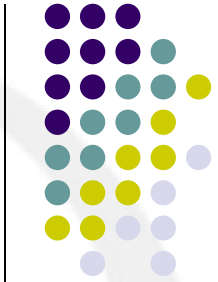
Regole associative – Subcategorie



- La regola
  - [USA E GETTA TAVOLA-STOV. CARTA COLORATA DECORATA] [USA E GETTA TAVOLA-ACCESSORI USA E GETTA] [USA E GETTA TAVOLA-STOVIGLIE PLASTICA BIANCA]→[USA E GETTA TAVOLA-TOVAGLIE-TOVAGLIOLI]
  - Support: 0,1 Confidence: 84,7 Lift: 12,767
- è supportata da 158 clienti. Il classificatore ottenuto ha una accuratezza del 88,13% su tutti i dati, e del 100% sui dati classificati come positivi.
- Profilo cliente:
  - Uomini celibi che lavorano per enti pubblici aventi 43-45 anni
  - Liberi professionisti aventi titolo di studio media inferiore di 59-60 anni
  - Militari di carriera sposati di aventi 38-41 anni

Evaluation – Obj 2

Regole associative – Subcategorie

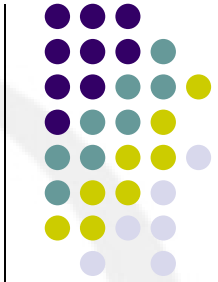


- La regola
  - [SNACK SALATI-POP CORN/CEREALI] [SNACK SALATI-ESTRUSI] [BIBITE-ARANCIATE] → [BIBITE-COLE]
  - Support: 0,1 Confidence: 82,2 Lift: 10,34
- è supportata da 155 clienti. Il classificatore ottenuto ha una accuratezza del 86,73% su tutti i dati, e del 100% sui dati classificati come positivi.

- Profilo cliente:
  - Ragazze disoccupate di 30-34 anni aventi un diploma magistrale
  - Vedovi di 57-60 anni liberi professionisti
  - Uomini/donne sposati di 32-40 anni e impiegati

Evaluation – Obj 2

Regole associative – Categorie

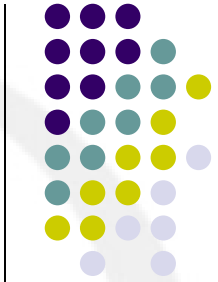


- La regola
  - [UOVA] [OF PREPARATA] [VERDURA FRESCA] [LATTE] [FRUTTA FRESCA] → [ORTAGGI]
  - Support: 0,8 Confidence: 85,2 Lift: 1,893
- è supportata da 1543 clienti. Il classificatore ottenuto ha una accuratezza del 68,61% su tutti i dati, e del 85,94% sui dati classificati come positivi.

- Profilo cliente:
  - Uomini/donne di 44-50 anni che lavorano nel privato e hanno studiato materie scientifiche
  - Donne single di 34-35 anni laureate che lavorano per enti pubblici
  - Donne sposate aventi 20-27 anni

Evaluation – Obj 2

Regole associative – Categorie



- La regola
  - [OLIO DI OLIVA] [UOVA] [SUINO] → [BOVINO]
  - Support: 0,7 Confidence: 78,9 Lift: 1,757
- è supportata da 1440 clienti. Il classificatore ottenuto ha una accuratezza del 64,83% su tutti i dati, e del 89,24% sui dati classificati come positivi.
- Profilo cliente:
  - Operai diplomati single aventi 33-34 anni
  - Dirigenti di azienda sposati aventi 39-70 anni
  - Casalinghe single diplomate aventi 33-39 anni

Evaluation – Obj 2

Pattern sequenziali – Articoli

- Il pattern sequenziale
  - { [PIZZAIOLA TRIS LOCAT.GR.125X3] } { [MOZZAR.S.LUCIA TRIS GR.125X3] }
  - Support: 0,010283
- è supportato da 599 clienti. Il classificatore ottenuto ha una accuratezza del 73,23% su tutti i dati, e del 94,16% sui dati classificati come positivi.
- Profilo cliente:
  - Artigiani aventi 29-30 anni
  - Single maschi e laureati aventi 60-70 anni
  - Donne laureate di 30-31 anni che lavorano per enti privati

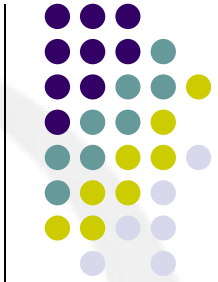
Evaluation – Obj 2

Pattern sequenziali – Articoli

- Il pattern sequenziale
  - { [IMPASTO PIZZA] } { [IMPASTO PIZZA] } { [IMPASTO PIZZA] }
  - Support: 0,007223
- è supportato da 426 clienti. Il classificatore ottenuto ha una accuratezza del 74,85% su tutti i dati, e del 99,59% sui dati classificati come positivi.
- Profilo cliente:
  - Donne di 29-54 anni appartenenti al clero
  - Uomini single di 39-40 anni che lavorano per enti pubblici
  - Donne sposate con laurea ma disoccupate di 33-35 anni

Evaluation – Obj 2

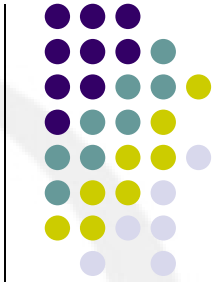
Pattern sequenziali – Subcategorie



- Il pattern sequenziale
  - { [AVICUNICOLO-POLLO] } { [BOVINO-VITELLONE] }
  - Support: 0,148801
- è supportato da 8650 clienti. Il classificatore ottenuto ha una accuratezza del 56,79% su tutti i dati, e del 73,06% sui dati classificati come positivi.
- Profilo cliente:
  - Donne diplomate casalinghe nubili di 32-36 anni
  - Uomini sposati di 72-76 anni che non lavorano
  - Impiegati sposati di 32-53 anni

Evaluation – Obj 2

Pattern sequenziali – Subcategorie

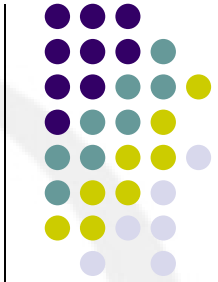


- Il pattern sequenziale
  - { [FORMAGGI A SERV.ASSISTITO-SEMIDURI/DURI] }
    { [FORMAGGI LIBERO SERVIZIO-FORMAGGI FRESCHI] }
  - Support: 0,100328

- è supportato da 5805 clienti. Il classificatore ottenuto ha una accuratezza del 53,58% su tutti i dati, e del 80,83% sui dati classificati come positivi.

- Profilo cliente:
  - Liberi professionisti sposati di 34-38 anni
  - Impiegati di 47-50 anni

Deployment -- Conclusioni



- Sono state analizzate le vendite di un ipermercato relative al primo trimestre del 2005 al fine di estrarre regole associative e pattern sequenziali che descrivono comportamenti di acquisto dei clienti.

- Sono stati inoltre determinati i profili dei clienti che supportano alcune delle regole associative/pattern sequenziali ritenute interessanti.

- Purtroppo non è stato possibile effettuare l'estrazione dei profili a tutti i livelli di astrazione, per via della generalità delle regole e pattern sequenziali ottenuti.

The End

Wake Up!!!