Politecnico di Torino

# Pre-Lab 2

## Business Intelligence per Big Data

*Estrazione di pattern frequenti e regole di associazione da dati strutturati e non strutturati*

| | | |
|---|---|---|
| FP-Tree | **Association rules** | **Outlier detection** |
| **Supporto,** confidenza**, correlazione** | **Binomial values** | Generalized association rules |

*Eleonora Poeta, Meryem Ennadi*

Obiettivo 1

# Analisi Esplorativa

*Capire i dati prima di estrarre pattern*

# Market Basket Analysis – Il Contesto

## Cos'è la Market Basket Analysis?

Tecnica di data mining per trovare associazioni tra elementi che co-occorrono spesso in insiemi (basket/transazioni).

**Esempi reali:**

- Prodotti comprati insieme in un supermercato
- Brani musicali nelle stesse playlist Spotify
- Pagine web visitate nella stessa sessione
- Sintomi che co-occorrono in diagnosi mediche

**Obiettivo:**
Scoprire pattern nascosti nei comportamenti degli utenti per supportare decisioni (raccomandazioni, layout negozio, bundling).

## Dataset: Playlists_Spotify.xlsx

**Struttura del file:**

- Righe = playlist degli utenti
- Colonne = brani musicali
- Celle = "true" se il brano è nella playlist, "false" altrimenti

**Domande esplorative:**

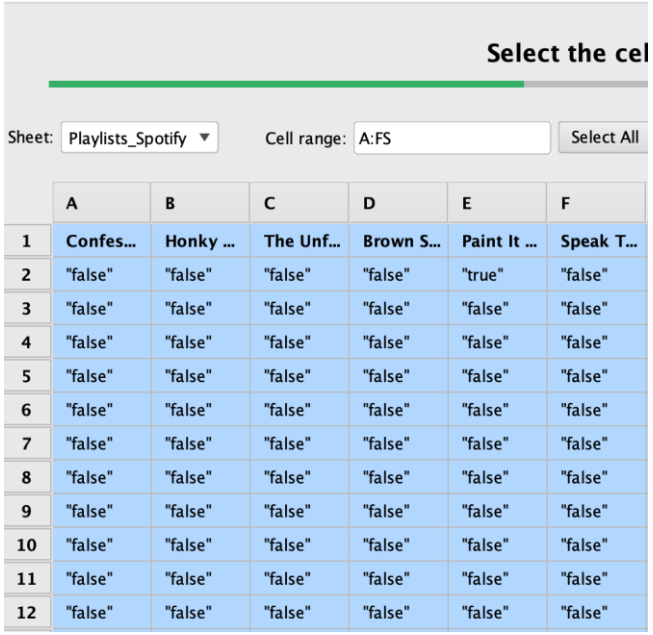
- Quante playlist sono disponibili? (→ n. righe)
- Quanti brani musicali? (→ n. colonne)
- Qual è il brano più popolare? (→ colonna con più "true")

Nel Lab: Read Excel → analizzare statistiche per rispondere alle 3 domande esplorative.

# Dataset utilizzati

**Playlists_Spotify.xlsx**

Questo dataset ontiene i nomi delle canzoni che fanno parte di almeno una playlist.
Le righe successive contengono le playlist degli utenti (ogni riga contiene una playlist).
In ogni riga la cella relativa a un brano musicale assume il valore “true” o “false” in funzione del fatto che in quella playlist ci sia oppure non ci sia il brano musicale a cui è associata la colonna.




| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Confes... | Honky ... | The Unf... | Brown S... | Paint It ... | Speak T... |
| 2 | "false" | "false" | "false" | "false" | "true" | "false" |
| 3 | "false" | "false" | "false" | "false" | "false" | "false" |
| 4 | "false" | "false" | "false" | "false" | "false" | "false" |
| 5 | "false" | "false" | "false" | "false" | "false" | "false" |
| 6 | "false" | "false" | "false" | "false" | "false" | "false" |
| 7 | "false" | "false" | "false" | "false" | "false" | "false" |
| 8 | "false" | "false" | "false" | "false" | "false" | "false" |
| 9 | "false" | "false" | "false" | "false" | "false" | "false" |
| 10 | "false" | "false" | "false" | "false" | "false" | "false" |
| 11 | "false" | "false" | "false" | "false" | "false" | "false" |
| 12 | "false" | "false" | "false" | "false" | "false" | "false" |

Nel Lab: Usare Read Excel + Import Configuration Wizard per caricare UsersSmall.xls e analizzare gli attributi.

Obiettivo 2

# Itemset Frequenti

*Analizzare dati strutturati in RapidMiner*

# Itemset Frequenti – Definizioni Chiave

| Item | Itemset | Transazione | Supporto |
| --- | --- | --- | --- |
| Un **elemento singolo** (es. un brano musicale, un prodotto). In Spotify: ogni brano è un item. | Un **insieme di uno o più item**. Es: {Never Say Never, All My Life}. | Una **singola riga** del dataset. In Spotify: una playlist di un utente. | **Frazione** di **transazioni** che **contengono l'itemset**. $sup(\{A,B\}) = |\{t: A,B \in t\}| / N$ |

## Supporto minimo (min_support) – Parametro chiave

Solo gli itemset con supporto ≥ min_support vengono considerati 'frequenti' e restituiti dall'algoritmo.
**Valore alto**: pochi itemset frequenti, solo i pattern più comuni. Rischio: perdere pattern rari ma interessanti.
**Valore basso**: molti itemset frequenti, rischio di combinatorial explosion e pattern non significativi.

📋 Nel Lab: FP-Growth con min_support=0.01 (1%). Parametri 'Min Size', 'Max Size', 'Contains Item' per esplorare i risultati.

# Analisi dei dati – Concetti Chiave

## Cos'è l'algoritmo FP-Growth?

L'algoritmo **FP-Growth** permette l'estrazione di pattern frequenti (come le combinazioni di brani nelle playlist) in maniera più efficiente di Apriori. Infatti, sfrutta:

- **Compressione dei Dati**: usa una struttura **FP-tree** (ad albero) per rappresentare il database in memoria.
- **Solo due scansioni** del database:
  - Una per contare i supporti degli item.
  - Una per costruire l'albero.
- **Ordinamento**: elementi ordinati in modo decrescente in base al loro supporto.
- **Visita Ricorsiva**: tramite cui estrae i pattern attraverso delle condizioni.

## Passi dell'algoritmo

**1** **Scansione 1**

Conta la frequenza di ogni item singolo. Elimina gli item sotto min_support.

**2** **Costruzione FP-Tree**

Inserisce le transazioni nell'albero ordinando gli item per frequenza decrescente.

**3** **Scansione 2**

Estrae i pattern frequenti percorrendo i percorsi condizionali dell'albero (divide et impera).

**4** **Output**

Lista di tutti gli itemset frequenti con il loro supporto calcolato.

Nel Lab: Operatore FP-Growth in RapidMiner. Input: matrice item-in-dummy-coded. Uscita 'fre' → itemset frequenti.

Obiettivo 3

# Regole di associazione

*Da identificare itemset frequenti a generare regole di associazione*

# Regole di Associazione– Concetti e Tecniche

## Cosa sono le Regole di Associazione?

- Sono pattern che esprimono la co-occorrenza frequente di elementi in un database transazionale.
- Una regola è definita nella forma A⇒B, dove A è il corpo (antecedente) e B è la testa (conseguente).
- Non implicano casualità, ma indicano che la presenza dell'elemento A è spesso accompagnata dalla presenza di B.

## Metriche di Valutazione

**Supporto** $sup(A \cup B) = |A \cup B| / N$

Indica la frazione di transazioni che contengono sia l'antecedente A che il conseguente B (AUB) .

**Confidenza** $conf(A \rightarrow B) = sup(A \cup B) / sup(A)$

Esprime la probabilità condizionata di trovare B data la presenza di A; ne indica l'*affidabilità*.

**Lift** $lift(A \rightarrow B) = conf(A \rightarrow B) / sup(B)$

Misura la correlazione tra A e B; se è maggiore di 1, la correlazione è positiva (gli elementi compaiono insieme più spesso di quanto ci si aspetterebbe per caso).

Nel Lab: Operatore "Create Association Rules" -> da collegare all'uscita del componente FP-Growth (impostare min confidence al valore 0.75)

# Regole di Associazione come Sistema di Raccomandazione

**Idea chiave: le regole A → B ci dicono cosa aggiungere a una playlist che contiene già A.**

Se la regola ha alta confidenza: gli utenti che hanno A nella playlist hanno B nel 75%+ dei casi → suggerisci B.

## Itemset frequenti → quando usarli

- Domanda: quante volte X e Y compaiono insieme?
- Utile per analisi esplorativa e frequenze
- Non hanno direzione (A con B = B con A)
- Esempio: trovare i brani più co-presenti nelle playlist

## Regole di associazione → quando usarle

- Domanda: dato X, cosa posso raccomandare?
- Hanno direzione (A→B ≠ B→A)
- Confidenza misura l'affidabilità del suggerimento
- Esempio: dato che hai 'All My Life', aggiungerei...

Nel Lab: Obiettivo 3 – estendere processo con Create Association Rules. Rispondere: a chi suggerire 'Never Say Never'?

Obiettivo 4

# Estrazione di associazioni da dati strutturati

*Analisi delle correlazioni  tra dati strutturati*

# Associazioni da dati strutturati – Perché e Come

## Perché estrarre associazioni da dati strutturati?

Consentono di indetificare delle *correlazioni* e forniscono *supporto alle decisioni*, consentono di comprendere le preferenze o i comportamenti di determinati profili di utenti.

### Gestione valori nulli

Identificazione del simbolo "?" come valore mancante e sostituzione con il valore più frequente tramite l'operatore "Replace Missing Values"

✓ Pro: preservazione dei dati(imputazione)

✗ Contro: rimuove record icompleti

### Discretization

Gli algoritmi di associazione lavorano su at tributi categorici quindi è fondamentale co nvertire variabili continue (Age) in variabili discrete.

✓ Pro: Interpretabilità

✗ Contro: riduzione del rumore

### Binomializzazione

Utilizzo dell'operatore "Nominal to Binomial" per trasformare ogni record in coppie "Attributo = Valore"

✓ Pro: normalizzazione

✗ Contro: dimensionalità

📋 Nel Lab: Tramite l'operatore Declare Missing Value, dichiarare per tutti gli attributi '?' come valore NULL e sostituirli con avg tramite "Replace Missing Values"

# Dati Strutturati: Pipeline di Preprocessing

### 1. Read Excel

Caricare UsersSmall.xls con Import Wizard.

### 2. Declare Missing

Dichiarare '?' come valore NULL.

### 3. Replace Missing

Sostituire NULL con il valore più frequente.

### 4. Discretize

Discretizzare Age (es. binning a 5 fasce).

### 5. Nominal to Binomial

Trasforma ogni attributo categoriale in colonne binarie: Attr=Valore → true/false.

### 6. FP-Growth + Rules

Estrarre itemset frequenti e regole di associazione.

📋 Nel Lab: Nominal to Binomial trasforma 'Education=Bachelors' → colonna binaria. FP-Growth input: 'items in dummy coded columns'.

# Interpretare le Regole – Lift e Conclusioni

**Una regola è interessante quando Lift > 1 (preferibilmente >> 1)**

Lift = 1 significa che A e B sono statisticamente indipendenti. La regola non aggiunge valore informativo.

| | |
|---|---|
| **Lift > 1** | A e B co-occorrono più del previsto → associazione positiva → regola utile. |
| **Lift = 1** | A e B sono indipendenti → la regola non è informativa. |
| **Lift < 1** | A e B co-occorrono meno del previsto → associazione negativa (anti-regola). |

**Analisi guidata:** Filtrare per conclusione (es. Response=Positive) → ordinare per Lift → identificare gli attributi predittivi più forti.

Obiettivo 5

# Estrazione di associazioni da dati testuali

*Analisi delle correlazioni da dati testuali: da documenti a regole tra termini*

# Associazioni da dati testuali

## Matrice Document-Term (Ripasso dal Lab 1)

Per analizzare la collezione ObamaNews, il testo non strutturato deve essere trasformato in un formato numerico leggibile dali algoritmi:

- Ponderazione TF-IDF: misura l'importanza di un termine rispetto all'intera collezione di documenti

.

- Term Frequency (TF): l'importanza aumenta proporzionalmente al numero di occorrenze del termine nel singolo documento.

- Inverse Document Frequency (IDF): penalizza I termini troppo comuni (es.articoli) per valorizzare quelli più rari.

**Operatore: Process Documents from Files**

## FP-Growth + Create Association Rules

### FP-Growth

Utilizzare loperatore 'FP-Growth per l'estrazione dei pattern frequenti (insiemi di elementi che compaiono spesso insieme nel testo).

### Create Association Rules

Utilizzare l'operatore "Create Association Rules" per estrarre le regole. Nella pratica prende questi pattern e genera le regole vere e proprie (nella forma "se succede A, allora è probabile che accada B")

Nel Lab: Generate Sales Data → Select Attributes (numerici) → Correlation Matrix → Normalize. Discutere quando è necessario normalizzare.

# Riepilogo del Lab

### ① Analisi Esplorativa

Read Excel → n.playlist, brani, popolarità

### ② Itemset Frequenti

Identificare eventuali pattern

### ③ Create Association Rule

Estrazione delle regole dai pattern -> confidence 0.75

### ④ Associazioni da dati strutturati

Nominal to Binomial → FP-Growth → lift, confidence

### ⑤ Associazioni da dati testuali

TF-IDF → Numerical to Binomial → regole tra termini

*Buon laboratorio!* 🚀