**Data Mining**

È una fase del processo di Knowledge discovery in database (KDD).

**KDD** è il processo non banale di identificazione di modelli validi, nuovi, potenzialmente utili e comprensibili nei dati.

* **Processo**: Composto da fasi. Non banale, deve comportare un tipo di ricerca o inferenza, non il calcolo di una quantità predefinita.
* **Valido**: I nuovi modelli devono essere applicabili a nuovi dati.
* **Nuovi**: I nuovi modelli non devono essere noti precedentemente.
* **Potenzialmente utili**: Devono portare ad azioni utili.
* **Comprensibili**: I modelli devono essere compresi da esseri umani, per facilitare la comprensione dei dati.

**Processo di KDD**:

1. Comprensione del dominio applicativo
2. Consolidamento dei dati
3. Selezione e pre-elaborazione
4. Scelta del compito di data mining
5. Scelta dell’algoritmo
6. Data Mining: Applicazione dell’algoritmo
7. Valutazione del pattern, ed eventuale ritorno al punto 1
8. Consolidamento della conoscenza: Documento da mostrare all’utente, o risoluzione di conflitti.

KDD è **iterativo e interattivo**, richiede l’intervento dell’analista per molte decisioni e per la scelta dei vincoli da imporre. La parte più importante del processo è la selezione e pre-elaborazione dei dati.

**Consolidamento in database**: Unire tutti i dati provenienti da fonti eterogenee.

**Selezione Pre-elaborazione** può includere: Selezione di un campione, Rimozione attributi correlati, combinazione attributi, Riduzione domini, Strategia per gestire valori nulli. Raggruppamento valori discreti.

**Trasformazione dati**: Normalizzazione valori.

**Codifica**: La rappresentazione deve essere adatta allo strumento di DM usato.

**Obiettivo del DM**: Scoperta di regole associative, che descrivono regolarità nei dati.

**Regole Associative**

Descrivono la correlazione tra eventi e possono essere regole probabilistiche.

Eventi correlati: quando sono osservati insieme frequentemente. Sono implicazioni nella forma A→B con A,B insiemi disgiunti.

**Problema di Scoperta**

Items I: insieme di letterali key:value (evento, outlook=rain). Itemset X: insieme di eventi. ({outlook=rain, play=n}). Transazione T: insieme di items che T incluso in I.

Supporto e Confidenza

Supporto(X): #Transazioni che contengono x / Totale transazioni

Supporto(X→Y) = Supporto(X∪Y).

I valori di supporto variano tra [0,1], 1= l’insieme x è in tutte le transazioni.

X→Y ha confidenza c in D, se tra tutte le transazioni che contengono x, esiste una frazione c che contiene anche y. conf(X→Y) = Supp(X∪Y)/Supp(x) = P(y|x).

I valori di confidenza variano tra [0,1], 1= perfetta associazione tra x, y.

Scoprire le regole associative: Scoprire tutte quelle con supporto e confidenza minimi, dove i valori minsup e minconf sono specificati dall’user.

**Due sotto-problemi**:

* Trovare gli insiemi che supporto superiore al minimo (insiemi frequenti): APriori.
* Generare tutte le regole con almeno la confidenza minima dall’insieme dei frequenti.

**APriori**

Se un insieme ha un supporto minore del minsup, tutti i suoi sottoinsiemi andranno al di sotto della soglia e possono essere scartati. È iterativo ed esamina il database. Inizia con gli insiemi di un elemento.

A ogni passo, costruisce gli insiemi di DIM k, usando gli insiemi di SIM k-1 e considerando i superset con un elemento in più. Quelli non frequenti in k-1 non vengono considerati.

Terminazione: Quando l’insieme dei k candidati è vuoto, e i più frequenti sono quelli per k-1.