Uno standard per il processo KDD

- Il modello **CRISP-DM** (Cross Industry Standard Process for Data Mining) è un prodotto neutrale definito da un consorzio di numerose società per la standardizzazione del processo di Knowledge Discovery in Databases
- E' costituito da 6 fasi:
  1. Comprensione del business
  2. Comprensione dei dati
  3. Preparazione dei dati
  4. Modellizzazione
  5. Valutazione
  6. Implementazione

Le fasi del CRISP-DM in dettaglio

- 1. Comprensione del business:
  l'attenzione è posta sugli obiettivi e requisiti del progetto da una prospettiva di business; viene definito il problema di data mining da risolvere

- 2. Comprensione dei dati:
  l'obiettivo è la raccolta dei dati e la formulazione di ipotesi

- *Queste prime due fasi rappresentano l'identificazione del processo di KDD che si vuole realizzare*

Le fasi del CRISP-DM in dettaglio

- 3. Preparazione dei dati:
  si individuano tabelle, record e variabili; i dati sono ripuliti in funzione degi strumenti prescelti per la modellazione
- 4. Modellazione:
  in questa fase vengono scelte ed applicate una o più tecniche di data mining
- 5. Valutazione:
  tramite l'analisi dei risultati si valuta se sono stati raggiunti gli obiettivi prefissati e si ipotizza una futura applicazione del modello
- *Queste tre fasi rappresentano l'approccio classico per la costruzione di un modello di data mining*

Le fasi del CRISP-DM in dettaglio

- 6. Implementazione:
  se il modello raggiunge gli obiettivi, si crea un piano d'azione per implementarlo ed utilizzarlo nel contesto applicativo

- *Questa fase, a valle della definizione e soluzione del problema di data mining vero e proprio, rappresenta la messa in produzione e sfruttamento del modello ottenuto; è il momento in cui l'investimento fatto inizia a pagare*

- Per maggiori informazioni, visitate il sito: www.crisp-dm.org

Un modello di processo KDD più generale

- Più in generale, un processo KDD può essere considerato l'applicazione del metodo scientifico al data mining
- Il metodo scientifico fu introdotto per la prima volta dal filosofo Francis Bacon nel *Novum Organum* del 1620
- Bacon descrisse il metodo scientifico come un processo in 4 fasi
  1. Definizione del problema da rislovere
  2. Formulazione di ipotesi
  3. Esecuzione di uno o più esperimenti per validare o rigettare le ipotesi
  4. Definizione e verifica delle relative conclusioni

Un modello di processo KDD più generale

La corrispondenza è definita come segue:

Metodo Scientifico

1. Definizione del problema da risolvere
2. Formulazione di ipotesi
3. Esecuzione di uno o più esperimenti per validare o rigettare le ipotesi
4. Definizione e verifica delle relative conclusioni

Processo di KDD

1. Identificazione obiettivi
2. Creazione dati target
3. Preprocessing dati
4. Trasformazione dati
5. Data mining
6. Interpretazione e validazione
7. Azione

Le 7 fasi del processo KDD

- **Fase 1: Identificare gli obiettivi**
- Definizione chiara di ciò che deve essere portato a termine, con finalità generali espresse in forma di obiettivi specifici
- Aspetti che dovrebbero essere tenuti in conto:
  - Definizione chiara del problema e di una lista di criteri per misurare i successi e i fallimenti
  - Scelta strumenti di data mining da utilizzare (dipende dal livello di spiegazione desiderato, tipologia di apprendimento impiegata o combinazione di entrambi gli aspetti)
  - Stima costo del progetto e stesura piano gestione risorse umane
  - Definizione data completamento/consegna del prodotto
  - Programmazione manutenzione sistema (es. aggiornamento modello quando sono disponibili nuovi dati)

Le 7 fasi del processo KDD

- **Fase 2: Creare i dati target**

- Individuazione delle sorgenti dati che verranno utilizzate nel progetto: DB relazionali, DW, file di testo, Weblogs…
- Analisi delle varie sorgenti (struttura, ridondanza, qualità…) e definizione di eventuali operazioni di trasformazione necessarie a rendere i dati omogenei e meglio utilizzabili dagli strumenti di dati mining

Le 7 fasi del processo KDD

- **Fase 3: Preprocessare i dati**

- Pulizia dei dati e risoluzione eventuali anomalie presenti:
  - Duplicazione valori
  - Presenza valori errati
  - Necessità di semplificazione (smoothing, eliminazione outliers)
  - Presenza valori mancanti

- Questa fase non è necessaria (già fatta con l'ETL) se i dati provengono da una sorgente di tipo Data Warehouse

Le 7 fasi del processo KDD

- **Fase 4: Trasformare i dati**

- Trasformazioni comunemente applicate per rendere i dati più trattabili dai metodi di data mining:
  - Normalizzazione (scalamento decimale, normalizzazione min-max, standardizzazione z-score, normalizzazione logaritmica…)
  - Conversione (trasformazione dati categorici in attributi numerici, discretizzazione dati numerici…)
  - Selezione delle variabili e delle osservazioni:
    alcuni algoritmi di data mining non possono lavorare con dataset troppo grandi e/o con dati contenenti troppe variabili e/o non sanno distinguere le variabili rilevanti da quelle irrilevanti

Fase 4: scelta delle variabili

- **Un possibile algoritmo:**
  - Date N variabili, si generi l'insieme S di tutte le possibili combinazioni
  - Si rimuova la prima combinazione C dall'insieme S e si generi il modello M utilizzando le variabili in C
  - Si misuri la bontà del modello M
  - Fintantoché S non è vuoto:
    - Si tolga una combinazione C' da S e sulla base di questa si costruisca un altro modello M'
    - Si confronti la bontà di M' con quella di M
    - Il migliore fra M ed M' sia definito M e sia utilizzato come nuovo termnine di paragone
  - Il modello M finale è quello scelto

Fase 4: scelta delle variabili

- Questo algoritmo, basandosi su una ricerca esaustiva, sicuramente determina la combinazione migliore di variabili
- Il problema è la sua complessità, dato che con n variabili a disposizione, il numero di possibili combinazioni è 2^n -1
- Generare e testare tutte le combinazioni per un dataset contenente molte variabili diventa quindi impraticabile
- Pertanto si utilizzano solitamente tecniche alternative:
  - Eliminazione di variabili
  - Creazione di nuove variabili più significative (es. rapporto, differenza, media, incremento/decremento percentuale fra 2 variabili…)

Fase 4: eliminazione di variabili

- Alcuni metodi hanno la capacità intrinseca di selezionare le variabili più interessanti (es. in base al guadagno informativo nella costruzione dei classificatori); per altri metodi invece le variabili meno significative vanno scartate a priori
- Variabili candidate ad essere eliminate:
  - Variabili di input fortemente correlate con altre sono ridondanti
  - Per dati categorici, qualsiasi variabile avente un valore v con un punteggio di previsione maggiore di una soglia prefissata (quasi sempre la variabile avrà il valore v, e non sarà pertanto d'aiuto nel differenziare il comportamento dei dati)
  - Se l'apprendimento è supervisionato, l'importanza di una variabile numerica può valutarsi in base alla media della classe e alla dev.std.
  - Possono essere determinate tramite apprendimento genetico

Le 7 fasi del processo KDD

- **Fase 5: Data mining**

- A seconda del risultato della valutazione (fase 6), l'intero processo è sperimentale e iterativo. Per questa fase occorre:
  - Scegliere il training set e il test set
  - Designare un insieme di variabili di input
  - Scegliere una o più variabili di output, se l'apprendimento è supervisionato
  - Scegliere i valori dei parametri per l'apprendimento
  - E finalmente… utilizzare lo strumento di data mining per costruire un modello generalizzato dei dati

Le 7 fasi del processo KDD

- **Fase 6: Interpretare e valutare**

- Scopo di questa fase è determinare la validità di un modello e la sua applicabilità a problemi esterni all'ambito di test; se i risultati sono accettabili la conoscenza acquisita viene trasformata in termini comprensibili agli utenti
- Per la valutazione si possono usare svariati metodi:
  - Analisi statistica
  - Analisi euristica
  - Analisi sperimentale
  - Analisi umana

Le 7 fasi del processo KDD

- **Fase 7: Agire**

- Consiste nel decidere come utilizzare in modo proficuo il modello che è stato ricavato; esempi di utilizzo sono:
  - Creazione di un rapporto tecnico (report) su ciò che è stato scoperto
  - Riallocazione o posizionamento abbinato di determinati prodotti
  - Invio di proposte commerciali o informazioni promozionali ad un campione selezionato di consumatori
  - Incorporazione di un modello sviluppato come sistema front-end per la scoperta di tentativi di frode, per la concessione di mutui/finanziamenti, per la valutazione della classe di rischio (medico, assicurativo, criminale…)
  - Predisposizione di un nuovo studio scientifico sulla base dei risultati