IL DATA MINING E IL RUOLO DEL DATA SCIENTIST

Il termine **Data Mining** è un termine inglese che si riferisce ai processi che consentono di estrarre informazioni valide da grandi volumi di dati, in formati comprensibili e potenzialmente utili. **Il Data Mining offre alle aziende la capacità di identificare in anticipo i modelli di comportamento e le tendenze del mercato**. Affinché ciò sia possibile, si cerca di stabilire relazioni logiche tra i dati raccolti anche in forma non strutturata, consentendo attraverso sistemi di analisi avanzati l'identificazione di questi modelli.

Il Data Mining è usato da aziende con un grande focus sui consumatori come vendita al dettaglio, finanza, comunicazione e marketing. Si stabiliscono relazioni tra fattori interni come prezzo, posizionamento del prodotto e fattori esterni come indicatori economici, concorrenza e demografia dei clienti, ciò consente di determinarne ad esempio l'impatto sulle vendite di nuovi prodotti, la soddisfazione del cliente e i profitti aziendali.

Queste informazioni hanno una importanza strategica: la loro corretta analisi e interpretazione consente di ottimizzare, prevedere e fare evolvere le strategie, includendo dati provenienti da diverse fonti che portano informazioni utili per un determinato processo, settore o business.

Grazie al data mining è possibile incanalare una grande quantità di dati in modo che siano processati per restituire analisi approfondite e fondamentali per i processi decisionali e di pianificazione.

I dati forniti vengono così messi in relazione strategica creando uno strumento analitico utilizzato ad esempio dalle aziende per aumentare le entrate e tagliare i costi.

Che cosa porta di nuovo il Data Mining?

Prima della sua creazione, esistevano già altri sistemi come i database ma la differenza è grande. I metodi antecedenti si limitavano a rispondere a questioni di riduzione della complessità mentre, nel caso di estrazione di dati, ricerche più complesse vengono eseguite al fine di individuare modelli e tendenze, anche per inferire regole. Il Data Mining può essere effettuato attraverso diverse metodologie, è l'applicazione di algoritmi specifici per estrarre modelli dai dati.

Il processo può essere riassunto in questi 5 passi:

1: selezione (da dati a dati target);

2: pre-elaborazione (dati target in dati elaborati);

3: trasformazione (elaborazione dei dati in dati trasformati);

4: modellizzazione (da dati trasformati a modelli);

5: interpretazione e valutazione dei modelli.

Questo processo richiede l'intervento di professionisti statistici che possano verificare il raggiungimento degli obiettivi proposti e la validità dei risultati ottenuti. Se esaminiamo come le modalità di analisi dei dati si sono evolute negli anni, notiamo che, a partire dagli anni '90, si è giunti all'utilizzo di una base dati creata appositamente: il **data warehouse**, che tuttavia consentiva soltanto una valutazione a consuntivo di quanto accaduto nel passato oppure di ciò che sta accadendo ora. Più recentemente, ha cominciato ad affermarsi la necessità di effettuare analisi previsionali, per anticipare gli eventi e ottenere un vantaggio. La grande quantità di dati disponibili ha reso necessaria l'adozione di tecniche di analisi efficienti ed in grado di lavorare su valori numerici, testuali o binari (per es. le immagini).

Le tecniche di analisi a cui ci riferiamo consentono di "scavare" nei dati ed estrarre informazioni, pattern e relazioni non immediatamente identificabili e non note a priori.

Il data mining può essere utilizzato in qualsiasi settore economico, industriale e non solo.

La componente tecnologica riveste una grande importanza, poiché gli algoritmi di data mining richiedono una certa potenza di calcolo; tecniche di ottimizzazione delle performance sono essenziali, soprattutto in presenza di una mole di dati elevata. Nel processo di data mining è però la **figura umana** ad assumere un **ruolo centrale**: si tratta, infatti, di un processo che richiede l'interazione di un esperto del business e di statistica, che deve sfruttare la propria conoscenza per la preparazione dei dati, per costruzione dei modelli e per la valutazione dei risultati.

È necessaria una buona conoscenza del business nel cui ambito si vogliono applicare le tecniche di data mining: ciò consente la corretta valutazione e selezione dei dati di partenza rilevanti. Occorre inoltre aver pienamente compreso i requisiti e gli obiettivi che si vogliono raggiungere, al fine di poter interpretare nel modo corretto i risultati dei modelli.

L'**applicazione del data mining a scopo predittivo** consente di determinare, in modo probabilistico, l'accadimento di eventi futuri, come per esempio il comportamento d'acquisto di un cliente, il grado di fedeltà della clientela, l'evoluzione della domanda di prodotti e servizi. Nel settore bancario, ad esempio, si può sfruttare il Data Mining per ricavare dei modelli predittivi sull'utilizzo delle carte di credito: possono essere utili per identificare le operazioni che hanno una probabilità più alta di essere fraudolente. Questo tipo di applicazione del Data Mining ha dunque lo scopo di individuare le probabilità del verificarsi di un determinato evento e di consentire di agire di conseguenza.

Buona parte degli algoritmi sono stati sviluppati in Nuova Zelanda come pure due strumenti molto conosciuti, WEKA ed R, il primo utilizzato anche nelle nostre università mentre il secondo più rivolto agli statistici perché con una curva di apprendimento piuttosto ripida. I più blasonati strumenti in commercio delle grandi società usano comunque gli stessi algoritmi di questi due (ed altri come Orange e Knime) liberi e gratuiti.

Il data mining può trattare dati qualitativi, dati quantitativi, dati testuali nonché immagini e suoni.

Oggi molto richiesta è la **figura del "Data Scientist"**, ma di che cosa si occupa?

- Progettare e interpretare gli esperimenti in chiave statistica per orientare le decisioni.
- Costruire modelli che prevedono il segnale, non il rumore, utilizzando la regressione, la classificazione, l'analisi delle serie temporali, es. analisi predittive dell'andamento di eventi nel tempo.
- Trasformare i grandi dati in un grande quadro d'insieme tramite il Clustering e la riduzione della dimensionalità, es. la segmentazione della clientela.
- Comprendere l'utenza e la clientela, la conservazione, la conversione e i contatti, tramite la regressione e l'analisi degli effetti causali, es. l'analisi dei clienti che possono passare alla concorrenza.
- Dare ai tuoi clienti e utenti quello che vogliono con la "Basket Analysis" e l'algoritmo Apriori.
- Stimare in modo intelligente con l'Analisi bayesiana dei dati, tecnica ben nota in passato e ritornata di grande attualità.
- Raccontare la storia con i dati, con opportune tecniche di visualizzazione.

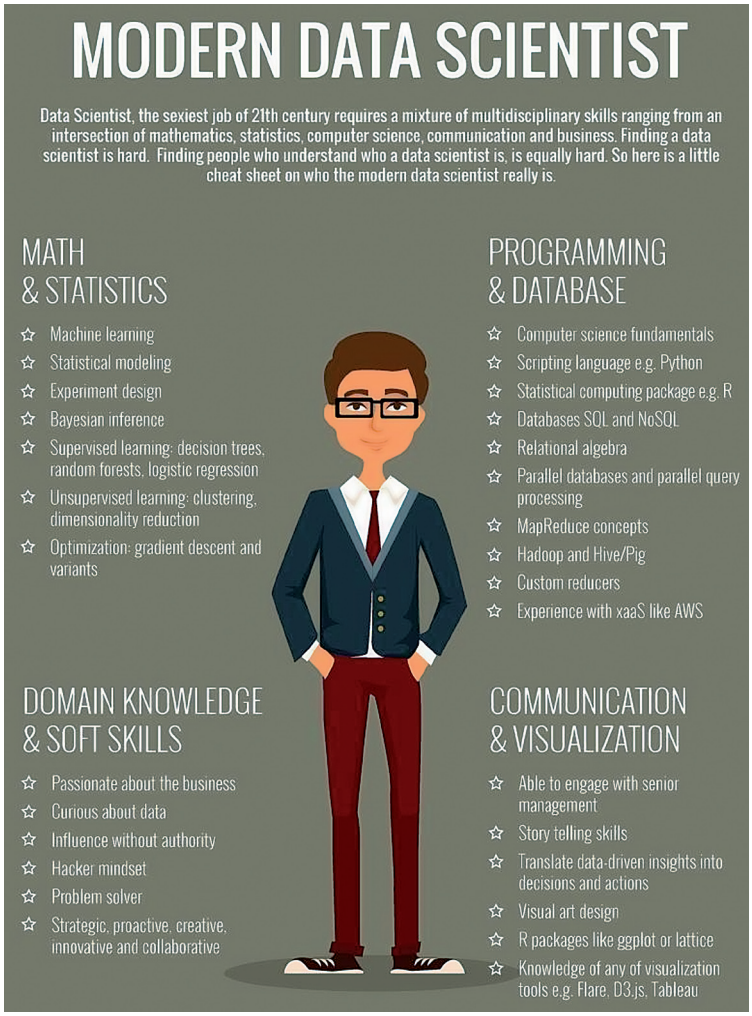

Immagine tratta da: https://image-store.slidesharecdn.com

Il **ruolo del Data Scientist in azienda è quello di ambasciatore tra i dati e l'azienda.**

La comunicazione è fondamentale e il Data Scientist deve essere in grado di spiegare le proprie intuizioni senza sacrificare la fedeltà dei dati. **Il Data Scientist non si limita a riassumere i numeri, ma spiega perché i numeri sono importanti e quali intuizioni attuabili si possono ottenere da questi**.

Il Data Scientist è un narratore, comunicando il significato dei dati e la loro importanza diffonde conoscenza, cosa essenziale in situazioni come le seguenti, nei campi più disparati: Analisi di Database (Estrazione di regole, Associazioni); Analisi di Mercato (Customer profiling, Marketing); Analisi di Rischio (Finance planning, Investimenti); Individuazione di Frodi (Carte di credito, Sofisticazioni alimentari); Supporto alle Decisioni (Resource management, Allocazione); Analisi Mediche (Diagnosi, Gestione donatori); Text mining (news-group, email, documenti) nel Web; Analisi di Politiche Economiche o Sociali (Rule learning); Analisi di Eventi Rari...