ANALISI DEI DATI PER LE IMPRESE INDUSTRIALI (2)
La preparazione dei dati

Nel primo articolo sul tema (pubblicato su Filo Diretto di giugno 2021) abbiamo affrontato in generale l'argomento dell'analisi; iniziamo ora a guardare un po' più in dettaglio gli aspetti specifici.

L'analisi dei dati, utilizzando opportuni algoritmi e scrivendo linee di programma, diventa poi ciò che viene chiamato "**Machine Learning**", che costruisce sistemi informatici, cercando di applicare le leggi fondamentali dei processi di apprendimento.

L'obiettivo dell'analisi dei dati è scoprire tendenze nascoste, modelli e relazioni.

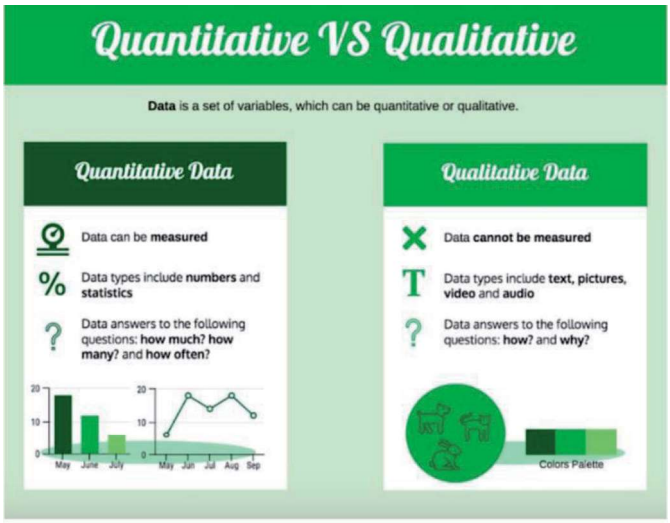
QUANTITÀ DEI DATI

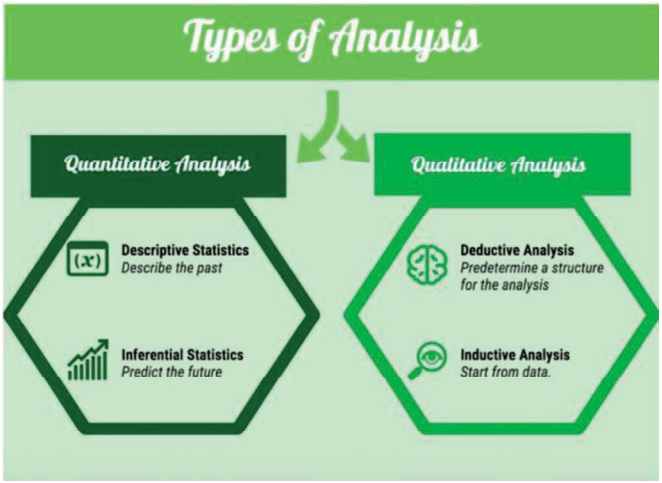
Fondamentale per eseguire delle analisi è avere dati in misura sufficiente. Ci basiamo sulla statistica ed è più efficace ottenere risultati da una certa mole che va poi ridotta, piuttosto che estrapolare conclusioni da poche informazioni.

TIPI DI DATI

I dati possono essere quantitativi o qualitativi. **Comprendere la differenza fra questi due è molto importante**, perché vengono trattati e analizzati in modi diversi.

I dati quantitativi includono quelli che possono essere espressi come numeri, che quindi possono essere misurati, contati e analizzati attraverso calcoli statistici. Possono essere utilizzati per descrivere e analizzare un fenomeno, al fine di scoprire tendenze, confrontare differenze ed eseguire previsioni. Spesso sono già strutturati, quindi è abbastanza facile eseguire ulteriori analisi.





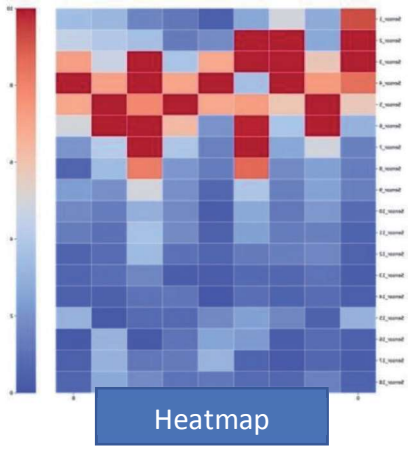
Includono sia i dati continui, che possono assumere qualsiasi valore numerico - come la temperatura - sia i dati discreti, che possono assumere solo determinati valori, come i numeri di scarpe.

I dati qualitativi non possono essere misurati attraverso tecniche di calcolo standard, perché esprimono sentimenti, sensazioni ed esperienze. Possono però essere utilizzati per comprendere nuovi aspetti di un fenomeno. Spesso i dati qualitativi non sono strutturati, quindi richiedono tecniche particolari per estrarre informazioni significative e necessitano di interpretazioni, ad esempio la "sentiment analysis".

ANALISI ESPLORATIVA

Exploratory Data Analysis (EDA), un termine coniato dal famoso statistico John Tukey, è una tecnica per comprendere e sviluppare una visione di insieme di dati prima di iniziarne l'indagine approfondita. Nell'EDA, le tecniche statistiche sono utilizzate per descrivere le caratteristiche dei dati quantitativi al fine di generare ipo-

tesi iniziali. Spesso i dati non sono in formati omogenei fra loro, ad esempio xlsx, csv, json…, il passo successivo è quindi quello di ricombinarli in un'unica forma.
Ci sono più modi per farlo: si può provare manualmente usando Ex-


Heatmap

cel ma uno dei modi più efficienti e veloci è quello di farlo con una libreria in *python*, ad esempio *pandas*.

ANALISI DI UN NUOVO SET DI DATI

Non appena si riceve un nuovo set di dati si vanno a ricercarne le proprietà e gli eventuali problemi, come dati mancanti o valori al di fuori dei campi attesi. **Si utilizzano strumenti di Statistica Descrittiva**, ovvero il calcolo di alcune misure di frequenza, di tendenza (media, mediana, moda) e di variabilità (massimo, minimo, varianza).
Si utilizza inoltre un'interpretazione qualitativa per comprendere il contesto dei dati stessi, per ricercare informazioni utili.

Una volta compresa la struttura tramite grafici di **distribuzione**, (immagine a destra) di **boxplot**,(si veda immagine a pag.29) che mettono in evidenza la mediana ed i valori estremi, e di **heatmap** (immagine a sinistra) per le correlazioni, si ragiona sulle azioni per razionalizzare il database. **Ogni distribuzione dei dati, dal punto di vista statistico, è diversa** e ha bisogno di un modo unico di pre-elaborazione.


Distribuzione

DATI MANCANTI

Se ci sono dati mancanti, si hanno diverse possibilità a disposizione, ad esempio sostituirli con la media e/o la mediana della serie oppure eliminarli o addirittura utilizzare alcune funzioni di Machine Learning per ipotizzarli. Esistono vari tipi di dati mancanti, ad esempio i **Missing Completely At Random (MCAR)**: non ci sarà alcuna relazione tra i dati mancanti e quelli osservati. In questo caso possiamo semplicemente rimuovere tutti i valori mancanti. Oppure i **Missing Data Not At Random (MNAR)**: ci sarà qualche relazione tra i dati mancanti e quelli osservati, dobbiamo gestire i valori mancanti sostituendoli con altri valori. Decidere se sostituire o addirittura cancellare i valori con dati mancanti è un processo lungo e complesso.

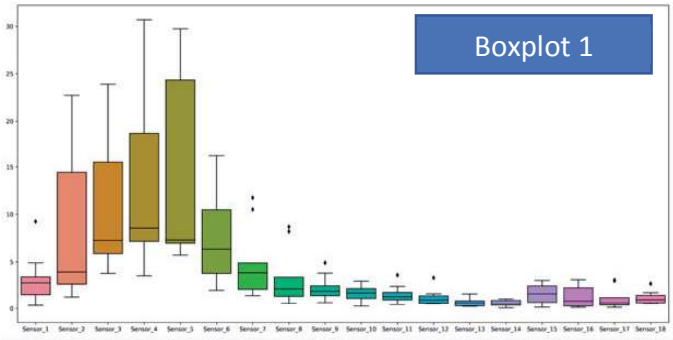
CORRELAZIONI

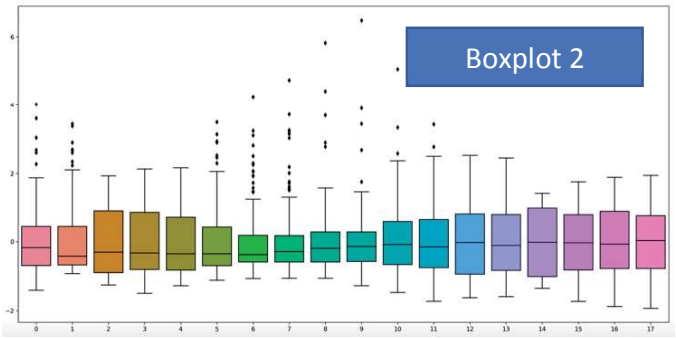
Per identificare le correlazioni fra i dati, si osserva se più serie di dati sono in qualche modo legate fra loro da una relazione. Questo influenzerebbe i successivi calcoli perché aumenta l'influenza di alcuni parametri rispetto ad altri. Si vedono bene graficamente con i colori (vedi figura Heatmap in alto a sinistra).

NORMALIZZAZIONE

A seconda di quali algoritmi utilizzeremo, potrebbe essere necessaria **un'operazione di normalizzazione o standardizzazione dei valori per ricondurli tutti allo stesso ordine di grandezza**. Questi processi sono richiesti per non dare più importanza ad una variabile rispetto ad un'altra, perché in questi casi la colonna con valori più grandi influenzerà maggiormente il risultato.
Ovviamente, se è di nostro interes-

se, possiamo fare il contrario, enfatizzarne una o più, anche combinando le variabili fra loro.
Vediamo gli effetti della standardizzazione nel grafico boxplot 2: rispetto boxplot 1, la media si parametrizza sullo zero e si riconoscono meglio i valori estremi (outlier).





OUTLIER

Sono dati insoliti che differiscono significativamente dal resto dei campioni e che possono facilmente influire sui risultati del modello. Possiamo usare due modi per gestirli: rimuoverli completamente oppure sostituirli con un valore adatto ai nostri scopi.

TIPI DI APPRENDIMENTO

Per effettuare l'analisi, il computer deve essere in grado di apprendere modelli. **Le due categorie più comuni di compiti sono l'apprendimento supervisionato e l'apprendimento non supervisionato.**

L'apprendimento supervisionato include attività per dati "etichettati", ovvero si dispone di dati obiettivo. Viene utilizzato spesso come modellazione predittiva, quindi per "indovinare" un certo risultato. Ogni osservazione deve essere etichettata con una "risposta corretta". Si può quindi costruire un modello predittivo visto che l'algoritmo sa che cosa è corretto, quindi è "supervisionato".
La **regressione** è l'attività supervisionata per creare modelli con variabili continue (quindi ad esempio prezzi, misure o grandezze numeriche), mentre la **classificazione** si usa per le variabili categoriche, ad esempio uomo-donna, caldo-freddo, funziona-non funziona….
L'apprendimento non supervisionato include attività per dati "non etichettati", ovvero non si conosce a priori una categorizzazione dei dati. In pratica viene spesso utilizzato come forma di analisi automatizzata dei dati per definire dei raggruppamenti con certe caratteristiche comuni. I dati senza etichetta non hanno una "risposta corretta" predeterminata, quindi si consente all'algoritmo di apprendere direttamente i modelli dai dati senza "supervisione".
Il "clustering" è l'attività di apprendimento senza supervisione più comune e serve per trovare gruppi e relazioni all'interno dei dati, ad esempio i clienti di un supermercato che acquistano certi prodotti.

QUALITÀ DEI DATI

L'elemento essenziale rimane la qualità dei dati. **Gli americani dicono: "Garbage In = Garbage Out", ovvero se i dati valgono poco, non si otterrà nulla, indipendentemente dagli algoritmi utilizzati.**
I "data scientist" di professione trascorrono la maggior parte del loro tempo ad analizzare e comprendere i dati, pulirli e prepararli prima di sottoporli agli algoritmi.
Nel prossimo articolo affronteremo i principali algoritmi utilizzati.

Immagini: a cura dell'autore e tratte da www.KDnuggets.com