Introduzione

Se avete questo libro tra le mani è probabile che un'idea – più o meno chiara – di cosa siano i Big Data Analytics ce l'abbiate già. Anche voi sarete stati travolti nelle vostre letture, durante gli studi o sul lavoro, da storie avvincenti e proclami categorici a supporto di una tesi che più o meno suona così: "Senza i Big Data non si va da nessuna parte". La sfida raccolta da questo volume è di farvi passare dagli aneddoti all'esperienza in prima persona, dalla teoria alla pratica, dalla frustrante consapevolezza delle potenzialità inespresse all'emozione di sporcarsi le mani e vedere le cose accadere. Il percorso che il libro vi propone è pensato per tre diverse categorie di lettori.

- Se siete manager di una qualsiasi funzione aziendale (dal marketing alle vendite, dalla finanza all'IT) e volete prendere decisioni consapevoli sull'utilizzo strategico dei tanti dati (a volte sembrano addirittura troppi) a disposizione nella vostra azienda, questo percorso fa per voi. Toccherete con mano gli algoritmi che stanno cambiando il modo di lavorare delle vostre organizzazioni e rimarrete stupiti da quanto voi stessi potrete trarre beneficio dal loro utilizzo. Tutto ciò vi renderà operatori autonomi delle tecniche di base della data analytics: i vostri colleghi apprezzeranno la vostra agilità nel parlare di questi strumenti, con la consapevolezza (e la soddisfazione) di chi li ha usati in prima persona.
- Se il vostro lavoro vi richiede di analizzare qualsiasi tipo di dati per effettuare scelte o per informare altre persone sullo stato del business, questo libro vi permetterà di diventare più efficienti (semplificando e automatizzando le vostre analisi) ed efficaci (adottando tecniche analitiche più sofisticate e sfruttando algoritmi di apprendimento automatico). In particolare, se siete già analisti o data scientist, i prossimi capitoli vi aiuteranno a riordinare le vostre conoscenze e a difendere con maggiore sicurezza il ruolo strategico del vostro ruolo nell'organizzazione di cui fate parte.
- Se, invece, state studiando o lavorate in un ambito anche molto diverso da quello della data science ma ne subite il fascino, questo libro vi aiuterà a muovere i primi passi nel mondo dei Big Data. Partendo da zero, acquisirete le tecniche base dell'analisi dei dati, iniziando un percorso di sviluppo che può portarvi a cambiare carriera o a crearvi – perché no – un'opportunità lavorativa come analista *freelance*.

Le tappe di questo ambizioso viaggio sono sette, una per ogni capitolo, e sono state pensate per rendere il vostro tragitto tanto graduale quanto avvincente.

1. Il primo capitolo metterà in ordine le idee, proponendo un modello concettuale per comprendere appieno (e saper spiegare a chi ci sta intorno) la vera novità portata dai Big Data e introducendo i termini che si incontrano più spesso. Questo è il capitolo più teorico di tutto il libro, ma ogni spunto concettuale, ogni esempio riportato e ogni tecnologia introdotta saranno sufficienti e necessari a capire appieno i capitoli successivi e a permettervi di distinguere i tanti "miti" dalla realtà.
2. Nel secondo capitolo introdurremo il nostro compagno di viaggio nella pratica dell'analisi dei dati e del machine learning, ovvero KNIME, un software che potete scaricare liberamente dal sito `www.knime.com` e installare sul vostro computer. Il capitolo introduce tutto ciò che è necessario per iniziare a usare KNIME in maniera da mantenere il controllo su tutti i passi analitici, senza essere obbligati a scrivere nemmeno una linea di codice.
3. Il terzo capitolo ci permetterà di acquisire agilità e velocità nel maneggiare dati. Impareremo come integrare e armonizzare flussi di dati diversi, ripulire e filtrare tabelle, importare ed esportare numeri in vari formati: tutte competenze imprescindibili per lavorare sui dati così come li ritroviamo nel mondo reale. In questo capitolo incontreremo il primo tutorial completo nel quale affronteremo, passo dopo passo, la costruzione di un processo analitico per interpretare le vendite di un business e-commerce.
4. Il quarto capitolo rappresenta il cuore metodologico di tutto il nostro viaggio, in quanto ci introdurrà nel mondo dell'intelligenza artificiale e ci farà capire cosa c'è dietro alla "magia" delle macchine in grado di imparare (*machine learning*). Il tutorial di questo capitolo ci permetterà di creare il nostro primo modello predittivo, in grado di prevedere il prezzo di vendita degli immobili a partire dalle loro caratteristiche.
5. Nel quinto capitolo impareremo ad addestrare le macchine a classificare automaticamente oggetti, utenti, transazioni o qualsiasi altra entità ci capiti sottomano. Vedremo come alcuni algoritmi di apprendimento automatico riescano a creare alberi decisionali in grado di svelare le caratteristiche – talvolta anche inaspettate e sorprendenti – delle varie classi di elementi analizzati. Il tutorial ci permetterà di capire come anticipare la propensione delle persone a compiere alcune scelte o effettuare particolari azioni attraverso un caso completo legato al business della telefonia mobile.
6. Il sesto capitolo spiega il modo in cui le macchine sono in grado di riconoscere strutture nascoste tra i dati, come similarità e differenze tra elementi di una tabella. Comprenderemo come sia possibile radunare persone, oggetti o eventi in gruppi omogenei tramite gli algoritmi di clustering e vedremo il tutto in azione nel tutorial che si propone di segmentare i consumatori per ottimizzare e personalizzare le campagne di marketing.
7. Nel settimo e ultimo capitolo completeremo il nostro viaggio attraverso una serie di approcci analitici più specifici (come il riconoscimento e l'analisi del linguaggio umano) e di estensioni di funzionalità di KNIME (come l'integrazione di R e Python o la connessione con piattaforme per l'analisi dei Big Data come Hadoop e Spark) che vi permetteranno di selezionare al meglio gli approfondimenti da affrontare dopo la fine del libro.

Nelle ultime pagine di questa introduzione è doveroso chiarire alcune delle scelte di stile e contenuto fatte allo scopo di rendere il vostro percorso quanto più efficace possibile.

- I tutorial sono stati pensati come parte integrante del percorso educativo proposto dal libro e non come approfondimenti o esercizi facoltativi. In effetti in ogni tutorial troverete il completamento dei concetti trattati nel capitolo, arricchiti da altri contenuti originali che, per ragioni didattiche, sono affrontati attraverso esempi concreti prima di essere messi a fuoco in maniera generale.
- Qualunque sia il vostro obiettivo di crescita e il vostro interesse, vi consiglio di eseguire integralmente i tutorial, passo dopo passo e "computer alla mano": offrono un'occasione unica per vedere in azione e fissare al meglio i concetti più centrali della data analytics e del machine learning, sfiorando anche gli aspetti più sfuggenti, difficilmente trasferibili se non attraverso la pratica e gli esempi concreti.
- Inoltre, nella loro ideazione, i tutorial sono stati pensati non solo per guidarvi nella soluzione dei casi specifici, ma anche per illustrarvi alcune tecniche di problem solving sviluppate da ogni data analyst con la pratica, attraverso l'esperienza nel mondo reale. Le continue revisioni di chi procede "per tentativi" nella costruzione di un modello, le necessarie interazioni con committenti e colleghi, le giuste domande da fare per capire come risolvere al meglio le esigenze di business: sono tutti "trucchi del mestiere" che i tutorial vi consentiranno di osservare.
- Per eseguire i tutorial avrete bisogno di ottenere i dati necessari al loro completamento: potete scaricare tutti i file menzionati nel testo dalla pagina dedicata a questo libro sul sito dell'editore (`www.apogeonline.com`) o su quello dell'autore (`www.aboutbigdata.net`).
- A chi tra di voi già lavora in questo ambito ed è colto dal dubbio su come sia possibile parlare di data science senza utilizzare esclusivamente Python o R rispondo con una sfacciata richiesta di fiducia: spero che, per la fine del libro, sia più chiara la scelta di usare KNIME per questo percorso. Vi anticipo che è basata soprattutto sull'esigenza di scorporare l'abilità di "saper fare analytics" da quella di "saper programmare". Inoltre, vedremo come KNIME permetta di integrare Python, R e altri linguaggi al suo interno: nessuna conoscenza pregressa, pertanto, verrà resa superflua e KNIME sarà uno strumento (spero apprezzato) da aggiungere alla vostra cassetta degli attrezzi.
- Nei capitoli che contengono i tutorial il formato dei numeri utilizzato è quello anglosassone e il separatore decimale è costituito dal punto invece che dalla virgola, come richiederebbe lo stile italiano: questo vi permetterà di ritrovare nel testo i numeri aventi lo stesso formato seguito da KNIME, semplificandovi la vita nell'affrontare i tutorial.
- Alcune delle immagini che incontrerete fanno uso dei colori per aumentarne la chiarezza didattica e semplificarne l'interpretazione. Per questo motivo il volume è stato arricchito da un inserto nel quale troverete una selezione delle immagini riportate nella loro versione a colori. Quando specificato nella didascalia, potrete trovare l'immagine nelle tavole a colori.
- Dobbiamo accettare il fatto che molte delle parole chiave nell'ambito della data analytics siano inglesi e che talvolta la traduzione in italiano rischi di essere una forzatura più che un aiuto all'apprendimento. L'uso delle parole in inglese nel testo è stato gestito in modo da accompagnarvi alla loro comprensione: nella maggioranza dei casi vengono spiegate al momento del loro primo utilizzo e mantenute nel loro formato originale nel resto del libro. Imparare a conoscere

e utilizzare questi termini vi permetterà di restare aggiornati dopo la lettura del libro, in quanto la forma inglese è quella più usata nell'ambito della data science, anche su testi scritti in italiano.

- Verso la fine del volume troverete un glossario di Big Data Analytics: è stato pensato come un "prontuario" da tenere a disposizione sulla scrivania, in modo da rintracciare all'occorrenza il significato dei termini più comuni.
- Ogni volta che ho dovuto effettuare una scelta di contenuto tra la completezza del rigore matematico e la semplicità di comprensione di un concetto ho preferito optare per quest'ultima via. Per questo motivo, nel libro troverete solo una manciata di formule matematiche malgrado l'argomento di cui parleremo sia intrinsecamente analitico. L'idea di fondo è quella di proporvi un modo "intuitivo" di afferrare (o rinforzare) l'essenza dei concetti di data analytics e machine learning, in modo da rendervi "operativi" e iniziare a usarli sul serio.

Dopo queste premesse non ci resta che tuffarci, con il primo capitolo, nelle potenzialità di Big Data Analytics. Buon viaggio!