Introduzione

Questo testo introduce ai fondamenti della data science con l'utilizzo dei due principali *linguaggi di programmazione* e *tecnologie open source*: R e Python, insieme ai rispettivi contesti applicativi formati da strumenti a supporto della scrittura di *script*, ovvero sequenze logiche di istruzioni con il fine di produrre determinati risultati o funzionalità, strumenti che possono essere di tipo CLI (Command Line Interface), *console* e terminali da usare con comandi testuali, e IDE (Integrated Development Environment) di tipo interattivo per il supporto all'uso dei linguaggi. Altri elementi che compongono il contesto applicativo sono le librerie supplementari contenenti le funzionalità aggiuntive oltre a quelle di base del linguaggio, package manager per la gestione automatizzata del download e installazione di nuove librerie, e poi documentazione online, cheatsheet, tutorial, fino ai forum online di discussione e aiuto per gli utilizzatori delle tecnologie e dei linguaggi. Tutto questo contesto, formato da un linguaggio, strumenti, funzionalità aggiuntive, discussioni tra utenti e documentazione online prodotta dagli sviluppatori, è quello che intendiamo quando diciamo «R» e «Python», non il semplice, mero strumento del linguaggio di programmazione che da solo sarebbe ben poca cosa. Come parlare solo del motore quando invece si vorrebbe parlare di come si guida un'automobile in strade trafficate.

R e Python, insieme e con l'accezione appena descritta, rappresentano il bagaglio di conoscenze per iniziare ad approcciarsi alla data science, svolgere i primi passi semplici, completare gli esempi didattici, prendere conoscenza con i dati reali, considerare funzionalità più avanzate, familiarizzare con altri dati reali, sperimentare casi particolari, analizzare meccanismi logici non evidenti, fare esperienza con dati reali più impegnativi, analizzare discussioni online su casi eccezionali, cercare fonti di dati nel mondo degli open data, ragionare sui risultati da ricavare, ancora più fonti di dati ora da mettere insieme, familiarizzare con formati di dati differenti, con grandi dataset, con dataset che vi faranno ammattire prima di riuscire a domarli, infine essere pronti per il salto ad altre tecnologie, altre applicazioni, usi, tipi di risultati, progetti di complessità sempre crescente. Questo è il percorso che si prospetta e, come si discuteva nella prefazione, è alla portata di chiunque ci si metta d'impegno, richiederà tempo e naturalmente un unico libro non può contenere tutto, ma può aiutare nell'avvio, mettere nella direzione giusta, accompagnare per un buon tratto.

Con questo testo partiremo dai passi elementari per prendere velocità rapidamente, utilizzeremo esempi didattici semplificati ma anche e soprattutto numerosi open data per cominciare immediatamente a familiarizzare con il tipo di dati che esistono nella realtà, invece che nella irrealtà degli esempi didattici. Termineremo affrontando esempi complessi, elaborati, nei quali anche le incoerenze e gli errori

che fanno parte della realtà quotidiana emergeranno e ci costringeranno a trovare delle soluzioni. Alla fine, guardandoci indietro, ci complimenteremo da soli per quanta strada saremo riusciti a percorrere da quei primi passi stentati.

1 Metodo di studio, metodo di lavoro

Spesso capita che studenti alle prese con questi contenuti, soprattutto i più giovani, si trovino inizialmente in difficoltà nel capire quale sia il modo giusto di affrontare lo studio e apprendere con efficacia. Una delle principali cause di questa difficoltà sta nel fatto che molti vengono abituati all'idea che l'obiettivo dell'apprendimento sia quello di non sbagliare mai e che da quello si misuri anche la qualità dell'apprendimento. D'altra parte, è in questo modo che funzionano quasi tutti gli esami, tanti più errori si commettono, tanto più basso sarà il voto. Non è questa la sede per discutere dell'efficacia delle metodologie d'esame o di filosofie didattiche, qui siamo pragmatici, l'obiettivo è imparare R e Python, la logica computazionale e quanto ci ruota intorno. Ma è proprio da una prospettiva del tutto pragmatica che sorge il problema dell'inadeguatezza dell'approccio che cerca di ridurre sempre e comunque al minimo gli errori e questo per almeno due buone ragioni. La prima è che inevitabilmente l'obiettivo di non sbagliare mai induce allo studio mnemonico, si memorizzano sequenze di passaggi, nomi, formule, frasi, casi specifici e si riduce la variabilità degli esempi che si considera, tendendo allo schematismo. La seconda ragione è semplicemente che cercare di non sbagliare mai è esattamente il contrario di quello che serve fare per imparare con efficacia R e Python e qualunque tecnologia digitale.

L'apprendimento di questi contenuti richiede, necessariamente, di svolgere molti esercizi pratici, rifare in modo meticoloso quelli proposti dal testo, ma anche variarli, introdurre modifiche, replicarli con dati differenti, tutti quelli degli esempi didattici possono ovviamente essere modificati, ma soprattutto anche tutti quelli con gli open data possono facilmente essere variati, anziché certe informazioni se ne usano altre, invece di un certo risultato se ne ricava uno diverso, oppure si usano dati diversi messi a disposizione dalla stessa fonte indicata per l'esempio. Insomma, per imparare bisogna lavorare anche con le mani, *hands-on* dicono gli americani, leggere soltanto non basta, leggere e memorizzare è la cosa più inutile. Bisogna procedere con metodo (essere metodici, meticolosi e pazienti sono requisiti fondamentali) e ricordare la seguente regola, perché di fondamentale importanza: *gli esercizi servono per sbagliare, un esercizio senza errori non serve a niente*.

2 Open data

L'utilizzo di open data reali, fin dai primi esempi e in misura largamente preponderante rispetto a esempi con dataset didattici semplificati, è una delle caratteristiche, forse la principale, di questo testo. Sono 24 i dataset tratti da open data utilizza-

ti, dei quali 9 italiani (da ISTAT, ministeri ed enti locali), 15 quelli internazionali di provenienza eterogenea, statunitensi, inglesi, tedeschi, svedesi, da grandi organizzazioni internazionali (Banca Mondiale, Nazioni Unite) così come da *charity* e istituti di ricerca indipendenti, osservatori sulle discriminazioni di genere, agenzie governative per il controllo del traffico aereo, la produzione e consumo di energia, l'emissione di inquinanti e altre informazioni ambientali, fino a dati resi disponibili da metropoli come Berlino e New York. Infine, immancabile, Wikipedia. Questa selezione non è che una goccia nel mare degli open data disponibili e in costante crescita, per quantità e qualità, anche italiani, per i quali si segnalano casi senz'altro virtuosi come, tra gli altri, i siti open data dei comuni di Bologna e Milano e, nonostante una certa farraginosità burocratica, i siti open data di alcuni ministeri, quello dell'Economia e delle Finanze in primo luogo. Una citazione a parte va fatta per ISTAT, l'Istituto Nazionale di Statistica, ed Eurostat, l'Ufficio di Statistica dell'Unione Europea, il cui contributo alla possibilità di realizzare ricerche, analisi, studi e anche didattica è impareggiabile.

Usare open data come si è fatto in questo testo è una scelta precisa che impone uno sforzo supplementare a chi intraprende il percorso di apprendimento, scelta basata sia sull'esperienza personale nell'insegnamento dei fondamenti della data science a studenti di scienze sociali e politiche (ogni anno ho anticipato sempre di più l'uso degli open data), sia sul difetto fondamentale di svolgere esempi ed esercizi con casi didattici, che sono sempre inevitabilmente irreali e irrealistici. Naturalmente i casi didattici, presenti anche in questo testo, si prestano perfettamente quando si tratta di mostrare una specifica funzionalità, un effetto o comportamento dello strumento computazionale ben preciso. Per quello il caso schematico e semplificato è ideale. Come già detto in precedenza, però, qui si tratta d'imparare a guidare nel traffico, non di limitarsi a comprendere alcuni meccanismi del motore, e l'unico modo per farlo è... guidare nel traffico, non c'è alternativa. Per noi è lo stesso, chiunque lavori con i dati sa che una delle competenze fondamentali è quella di saper preparare i dati per poter essere analizzati e da quelli ricavare i risultati (prima ci sarebbe quella di trovarli i dati, cosa altrettanto fondamentale) e sa anche che questo compito può facilmente essere la parte più gravosa, per tempo e sforzi, di tutto il lavoro. Studiare prevalentemente con esempi didattici semplificati cancella questa parte fondamentale di conoscenza ed esperienza, per questo motivo sono sempre irreali e irrealistici, comunque li si cerchi di aggiustare. Non c'è alternativa a mettere le mani e sbattere la testa su dati reali, maneggiare dataset anche di centinaia di migliaia o milioni di righe (quello di maggiori dimensioni che usiamo ampiamente in questo testo ha più di 500000 righe, sono i dati di tutti i voli interni degli Stati Uniti del mese di gennaio 2022) con i loro errori, le loro inconsistenze, le spiegazioni che si devono leggere e talvolta si interpretano male, perfino con casi in cui sono stati registrati dati in maniera incoerente (vedremo un caso bizzarro). Vanno conosciuti il prima possibile, per familiarizzare con il contesto reale e toccare con mano il fatto che dietro ai dati che possiamo e vogliamo usare ci sono organizzazioni composte da persone, non macchine o algoritmi, ed è grazie a loro se disponia-

mo dei dati dai quali estrarre nuove informazioni e conoscenza, bisogna armarsi di pazienza e districare ogni nodo, un passo alla volta. Questo fa parte dei fondamenti da apprendere.

3 Cosa non si impara

Un libro solo non può coprire tutto, lo abbiamo già detto ed è scontato, ma il punto da decidere è cosa lasciare fuori. Una possibilità è che l'autore cerchi di presentare quanti più argomenti diversi gli vengono in mente, questo è il modello enciclopedico, di lunga tradizione ma poco compatibile con un numero ragionevolmente limitato di pagine. Non a caso le più celebri tra le enciclopedie sono pubblicazioni da decine di ponderosi tomi. La versione breve del modello enciclopedico è una «sintesi», cioè una panoramica necessariamente non molto approfondita su svariati temi complessi. Molti testi didattici scelgono questa forma, che ha il vantaggio dell'ampiezza di temi toccati e permette di semplificare molto.

Questo libro ha una forma ibrida. È più ampio della norma perché include due linguaggi, invece di uno, ma non ha la forma della sintesi perché si concentra su un certo tipo specifico di dati e di funzionalità: i data frame, con l'aggiunta finale delle liste/dizionari, le operazioni di trasformazione e di pivoting, di indicizzazione per gruppi, aggregazione, trasformazioni avanzate e join di data frame, e su questi si scende nei dettagli (questo insieme di operazioni prende il nome di *data wrangling*). In breve, ci si occupa dei ferri del mestiere, indispensabili per fare data science.

Cosa è rimasto fuori? Moltissimo, ovviamente. Le tecniche e gli strumenti per *data visualization* (grafici), la parte relativa a modelli, sia descrittivi che predittivi, incluse le tecniche di machine learning, ovviamente la parte di analisi statistica (anche se questa è parte tradizionalmente a se stante), le tecnologie per cosiddetti «Big Data», cioè infrastrutture software distribuite e scalabili in grado di gestire non solo molti dati ma soprattutto flussi di dati, le molte estensioni web-oriented a partire da tecniche di raccolta di dati fino all'integrazione con dashboard e servizi web. Ancora, esistono standard specialistici, come per i dati climatici, dati finanziari, dati biomedici e codifiche utilizzate da alcune delle grandi istituzioni internazionali che non vengono trattati. L'elenco potrebbe proseguire.

Tutte queste ulteriori conoscenze, che fanno parte della data science, meritano di essere apprese. Per farlo occorrono gli strumenti di base che questo libro presenta. Oltre a questi giocano inevitabilmente gli interessi e i percorsi culturali e professionali di ognuno che potranno spingere in una certa direzione o in un'altra. Però di nuovo, una volta che si è verificato in prima persona che è possibile, qualunque background si abbia, acquisire con profitto i fondamenti della disciplina con R e Python, tutti gli ulteriori approfondimenti e sviluppi possono essere affrontati, esattamente con lo stesso approccio e spirito con i quali si sono imparati i fondamenti.