Società Italiana di Statistica
Istituita come ente morale con R.D. 13 luglio 1939

Parere sull'opportunità di creare una nuova classe di laurea magistrale in DATA SCIENCE

Audizione CUN, Roma, 14 giugno 2018

Premessa - La Statistica è anche scienza dei dati in continua evoluzione.

La Statistica è una disciplina scientifica che studia i metodi di raccolta, di organizzazione, di analisi, d'interpretazione e presentazione dei dati, e tutti le riconoscono un ruolo preminente nel progresso della conoscenza in qualsiasi campo del sapere. Già da oltre un secolo, ha sviluppato anche metodologie per gestire ed analizzare grandi moli di dati, come quelli rilevati in occasione dei Censimenti e di indagini campionarie su vasta scala o presenti in archivi amministrativi, definendo i principi che stabiliscono l'attendibilità dei dati raccolti (profilo di errore) e l'attendibilità dei risultati delle analisi (https://www.sis-statistica.it/ita/3445/Documenti%20della%20SIS).
Lo sviluppo dei metodi per l'analisi statistica e demografica è sempre stato affiancato dalle tecnologie di memorizzazione e di calcolo più moderne per l'epoca, che restituissero analisi efficaci e complesse dei dati trattati: integrazione di dati e di fonti, fusione di dati, *data mining*, analisi dei dati funzionali, analisi multivariate, simulazioni, algoritmi di stima non sarebbero stati possibili senza un continuo dialogo e scambio con gli studiosi di Matematica e di Informatica. Tanto che già nella seconda metà del secolo scorso si parlava di *data analysis* e *data driven analysis*, di *learning by data*, di *data science* e ci si interrogava sulla validità delle conoscenze raggiunte.

Attualmente, le classi di laurea di Statistica triennali e magistrali già consentono di disegnare dei percorsi formativi specifici in grado di formare professionisti con marcate abilità nella gestione e analisi di basi dati di grandi dimensioni che sono ampiamente diffusi in molti contesti lavorativi (ad esempio in contesti economici, industriali, informatici, medici, ecc.). Particolare attenzione è anche rivolta all'aspetto della comunicazione dei risultati e della visualizzazione dei dati.

Qualora si voglia comunque identificare una nuova classe di laurea che si ponga l'obiettivo di rispondere specificatamente al fabbisogno formativo del *Data Scientist*, è necessario che contenuti imprescindibili siano rappresentati dai metodi di valutazione della qualità dei dati, dalla probabilità, dalla modellistica statistica, dallo *statistical reasoning* e *learning* e che quindi possano essere dedicati all'ambito statistico probabilistico un numero congruo e prefissato di crediti.

Obiettivi culturali della classe, contenuti disciplinari e competenze trasversali indispensabili

Il principale obiettivo culturale della nuova classe è la formazione dello "Scienziato dei dati", che sappia riconoscere e gestire un approccio scientifico ibrido, in cui la modellizzazione *top-down* dei fenomeni trovi una nuova sintesi con la scoperta di conoscenze *bottom-up*, *data-driven* e spesso frutto di applicazione di algoritmi che generano le grandi masse di dati disponibili, applicando il ragionamento statistico induttivo e deduttivo a grandi moli di dati.

A tal fine è necessario distinguere il profilo di *Data Analyst* da quello di *Data Engineer*.

Nel primo caso - *Data Analyst* - è fondamentale sviluppare abilità analitiche, quindi propensione per il ragionamento matematico e statistico e sensibilità alla validazione e alla qualità dei dati, competenze di programmazione e anche doti comunicative, utili per presentare i risultati dell'analisi di dati complessi in forma chiara e comprensibile. Nel secondo, - *Data Engineer* - le abilità tecnico informatiche legate alla raccolta, sistemazione e mantenimento dei dati sono prevalenti, per garantire la disponibilità, la qualità e la fruibilità dei dati per l'analisi. Di non secondaria importanza è anche la competenza di *Data Protection* per la quale sono importanti, oltre alle competenze informatiche, anche quelle giuridiche, etiche e di *statistical disclosure* per il mantenimento e la tutela della *privacy*.

I laureati nei corsi di laurea magistrale della classe Data Science devono:

- saper coniugare i metodi e le tecniche della statistica, della matematica e della ricerca operativa con le tecnologie e metodologie dell'informatica, possedendo competenze in ciascuna delle aree;

- saper operare in gruppi interdisciplinari costituiti da esperti con competenze negli ambiti della matematica, delle tecnologie dell'informatica e della statistica, nonché con competenze proprie di specifici contesti applicativi con un'elevata capacità ad affrontare e risolvere i problemi (*problem solving*);
- saper affrontare problematiche connesse con l'utilizzo delle tecnologie informatiche e telematiche (con riferimento, tra gli altri, ai problemi di sicurezza e di tutela della riservatezza);
- saper comunicare efficacemente i risultati delle analisi condotte, in forma scritta e orale, anche per mezzo di tecniche di visualizzazione e rappresentazione delle informazioni avanzate con forte potenziale comunicativo (*data visualization*);
- essere in grado di utilizzare fluentemente, in forma scritta e orale, almeno l'inglese, con riferimento anche ai lessici disciplinari. In uscita, al termine del corso di studio magistrale, la conoscenza delle lingue straniere deve essere equiparabile almeno al livello B2.
- conoscere le problematiche dei fenomeni relativi ad uno o più contesti applicativi (economico, sociale, sanitario, demografico, biomedico, ambientale, tecnologico, ecc.) e le relative specificità metodologiche;

Dalle discussioni anche internazionali emerge che gli ambiti Matematico, Statistico e Informatico sono fondamentali per la formazione in *Data Science*. Tutte e tre dovrebbero essere presenti in un corso di laurea magistrale con un numero di crediti dipendenti dall'orientamento del corso di studio, in modo da consentire caratterizzazioni diverse. Gli argomenti da trattare nei corsi, per permettere conoscenze che portino alle abilità prima elencate, sono riassunti nel prospetto seguente per ciascuno degli ambiti citati ed anche per gli eventuali aggiuntivi ambiti applicativi di interesse:

Ambito Statistico	Ambito Matematico	Ambito Informatico	Ambito Sostantivo
inferenza statistica, analisi statistica multivariata, modelli statistici avanzati, statistica computazionale, *statistical learning*.	metodi di ottimizzazione, calcolo numerico, calcolo delle probabilità	calcolo ad alte prestazioni, calcolo distribuito e su cloud, trattamento di dati in forma di testo e immagini, data mining, machine learning e deep learning	conoscenze in settori applicativi al fine di ottenere analisi tematiche approfondite su specifici temi (biologici, ambientali, medici, sociali, economici, tecnologici, ecc.)

Le abilità sopra richiamate non possono essere acquisite senza una solida conoscenza di alcune discipline di metodo. Queste potranno essere individuate stabilendo un numero minimo di CFU in alcuni settori. Negli ambiti caratterizzanti dell'ordinamento saranno contenuti i settori delle discipline prima richiamati, in modo che sia possibile la definizione di percorsi in grado di omogeneizzare le necessarie conoscenze di base (es: ambito statistico, ambito informatico, ambito matematico). Mentre gli ambiti affini garantiranno la multidisciplinarità delle conoscenze e gli approfondimenti specifici con attività formative finalizzate all'acquisizione di competenze di alto livello nei campi applicativi di interesse.
Condividiamo un orientamento verso una ripartizione equilibrata dei CFU per attività obbligatorie tra le tre aree disciplinari fondanti (Informatica, Matematica e Statistica).

Naturali sbocchi professionali, o sbocchi verso il proseguimento degli studi

Nell'elenco che segue sono presenti ambiti di lavoro e possibili professioni coerenti con gli obiettivi della classe:
consulenza in senso lato, per aziende e pubblica amministrazione, centri di ricerca in ambiti diversi (biomedico, chimico, fisico, economico-sociale), centri di produzione di dati, statistiche pubbliche e ufficiali, analisi dei mercati finanziari, banche e imprese di assicurazione, *business analytics* e ricerca operativa, qualsiasi altro ambito in cui emergano big data utilizzabili a supporto del processo di decisione.
Il proseguimento degli studi potrà avvenire nei master di secondo livello e nei corsi di dottorato in Data Science (diversamente declinato) o in altre discipline.

Necessità di introdurre altri elementi (tirocini o stage, attività laboratoriali, competenze linguistiche, eccetera) indispensabili per il raggiungimento degli obiettivi della classe.

Le attività di laboratorio potrebbero essere previste soprattutto al fine di sviluppare la capacità di lavorare su progetti (definizione degli obiettivi, scelta dei dati e dei metodi, implementazione, interpretazione, presentazione). Anche alcune attività esterne, come tirocini formativi, presso enti o istituti di ricerca, laboratori, aziende e amministrazioni pubbliche, oltre a soggiorni di studio presso altre università italiane ed europee costituirebbero un utile strumento in tal senso. La conoscenza della lingua inglese potrebbe essere efficacemente sviluppata impartendo parte degli insegnamenti in lingua inglese.