Denominazione Insegnamento/Altra Attività	**FONDAMENTI DI DATA SCIENCE E MACHINE LEARNING**
SSD	INF/01
CFU	9
Ore	72
Tipologia Attività Formativa - Ambito	Obbligatorie – Discipline Informatiche
Obiettivi Formativi: Risultati d'Apprendimento Previsti e Competenze da Acquisire (Descrittori di Dublino)	L'insegnamento mira a fornire le competenze metodologiche e tecnologiche necessarie per estrarre conoscenza da grossi volumi di dati, mediante tecniche di data profiling, data mining e machine learning, utilizzando opportune strategie di visualizzazione dei risultati. In particolare, si intendono integrare le conoscenze di data management acquisite nell'ambito di altri corsi di basi di dati con competenze utili allo svolgimento della professione del data scientist. **Conoscenza e Capacità di Comprensione** Le Principali Conoscenze Acquisite Saranno: • Big Data • Data wrangling • Estrazione automatica di dipendenze tra i dati • Data quality e data cleansing • Data integration • Data mining • Mapreduce • Funzioni di similarità • Machine learning • Reti neurali **Capacità di Applicare Conoscenza e Comprensione** Gli studenti acquisiranno la capacità di: • Acquisire, organizzare, gestire ed elaborare grandi volumi di dati • Estrarre conoscenza dai dati • Selezionare dati utili • Organizzare un progetto basato su tecniche di machine learning • Comunicare la conoscenza estratta dai dati attraverso diverse forme di rappresentazione, incluso tecniche visuali.
Prerequisiti	Lo studente deve conoscere i fondamenti di data management, sistemi distribuiti, paradigma ad oggetti ed un linguaggio di programmazione.

Contenuti/Programma	Dopo una panoramica sui nuovi scenari applicativi legati alla gestione di grandi collezioni di dati distribuiti ed eterogenei, incluso le potenzialità di tecnologie capaci di estrarre conoscenza dai dati, il corso si concentrerà sui seguenti argomenti: **Big data (4 ore di teoria)** • Problematiche di big data (2 ore di teoria) • Tecnologie di supporto ai big data (2 ore di teoria) **Data preparation (12 ore di teoria)** • Data profiling (4 ore di teoria) • Dipendenze funzionali approssimate e loro utilizzo nel data quality (4 ore di teoria) • Integrazione dati da sorgenti multiple (4 ore di teoria) **Estrazione di conoscenza da grandi collezioni di dati (12 ore di teoria)** • Mapreduce (4 ore di teoria) • Valutazione della similarità (5 ore di teoria) • Introduzione al data mining (2 ore di teoria) • Algoritmo apriori (1 ora di teoria) **Machine learning (24 ore di teoria)** • Concetti introduttivi (4 ore di teoria) • Fasi di un progetto di machine learning (5 ore di teoria) • Modelli di Addestramento (2 ore di teoria) • Classificazione/regressione (3 ore di teoria) • Alberi di decisione (2 ore di teoria) • Ensemble learning and random forest (2 ore di teoria) • Clustering (2 ore di teoria) • Riduzione della dimensionalità (2 ore di teoria) • Support vector machine (2 ore di teoria) **Reti Neurali (14 ore di teoria)** • Introduzione alle reti neurali (2 ora di teoria) • Tensor flow (2 ore di teoria) • Percettroni Multilivello e Reti Neurali Profonde (2 ore di teoria) • Reti convoluzionali (2 ore di teoria) • Reti ricorrenti (4 ore di teoria) • Autoencoder (2 ore di teoria) **Strumenti per la data science (6 ore di teoria)** • Il linguaggio python (4 ore frontali) • Weka (2 ore frontali)
Metodi Didattici	L'insegnamento prevede 66 ore di didattica frontale su argomenti teorici e 6 ore su linguaggi e strumenti applicativi, con l'obiettivo di presentare i concetti e sviluppare capacità di progettare e implementare soluzioni per problematiche di data science e machine learning. Gli argomenti del programma vengono presentati con l'ausilio di presentazioni powerpoint, stimolando discussioni critiche con la classe. Per ogni argomento trattato, vengono illustrati possibili task che possono essere oggetto di un progetto di corso da parte di uno o più studenti. Per quanto riguarda gli strumenti applicativi, oltre all'utilizzo di

	presentazioni powerpoint, nelle quali vengono presentati concetti ed eventuali link a forum, manuali e siti di approfondimento, durante l'orario di ricevimento viene offerta agli studenti la possibilità di chiedere supporto in merito a simulazioni da essi effettuate sul proprio computer, di chiedere chiarimenti e risolvere eventuali problemi tecnici insieme al docente.
Modalità di Verifica dell'Apprendimento	Il raggiungimento degli obiettivi dell'insegnamento è certificato mediante il superamento di un esame con valutazione in trentesimi. L'esame prevede una prova scritta (in alternativa, una prova in itinere a metà corso) ed una prova orale. Inoltre, opzionalmente, gli studenti possono sviluppare un progetto per incrementare il voto ottenuto con le suddette prove. La prova scritta (o quella in itinere) mira ad accertare l'acquisizione dei concetti teorici. La prova orale invece consiste in un colloquio con domande e discussione sui contenuti teorici e metodologici trattati a lezione ed è finalizzata ad accertare la capacità di conoscenza e comprensione, nonché la capacità di esposizione dei concetti. La prova orale rappresenta la prova finale, pertanto essa può essere svolta solo dopo il superamento della prova scritta e, qualora si sia optato anche per lo sviluppo del progetto, anche dopo il completamento e la discussione di quest'ultimo. Il progetto è finalizzato ad accertare la capacità di applicare le conoscenze acquisite in scenari reali. Esso può essere svolto individualmente o in gruppi di massimo 3 persone, scegliendo tra un ventaglio di proposte fatte dal docente. Durante lo svolgimento del progetto, gli studenti potranno interagire con il docente al fine di comunicare gli stati di avanzamento dello stesso e le eventuali criticità emerse, concordando obiettivi e modalità di prosecuzione. Al termine del progetto, gli studenti devono consegnare al docente una tesina contenente la documentazione di progetto ricevendo, dopo qualche settimana, una prima valutazione dello stesso. Quest'ultima potrebbe contenere richieste di integrazione e/o di revisione del lavoro svolto. Pertanto, occorre sottomettere i risultati del progetto con diverse settimane di anticipo rispetto alla data in cui si intende sostenere la prova orale, onde consentire al docente di effettuare la correzione ed al gruppo di progetto di apportare le eventuali correzioni richieste. Al termine del progetto, al gruppo di progetto potrebbe essere richiesto di preparare una tesina ed una presentazione powerpoint della durata di circa 30 minuti. Il voto finale scaturisce, generalmente, dalla media dei voti in trentesimi conseguiti alla prova scritta (alternativamente, quella in itinere) ed a quella orale, con la possibilità di incrementare il punteggio così ottenuto fino a 3 punti, tramite lo sviluppo del progetto.
Testi di Riferimento	1. Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, Mining Of Massive Datasets", 3^ Edizione, Cambridge University Press, 2020. 2. Aurélien Géron, " Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems ",3^ Edizione, O Reilly ed, 2023. 3. Chirag Shah, A Hands-On Introduction To Data Science, Cambridge University Press, 2020.

	4. Foster Provost, Tom Fawcett, Data Science For Business: What You Need To Know About Data Mining And Data-Analytic Thinking, O Reilly Ed, 2013. 5. P. Deitel, H. Deitel, Introduzione A Python – Per L'informatica E La Data Science, Pearson 2021.
Altre Informazioni	La frequenza del corso è fortemente consigliata. Gli studenti devono essere preparati a trascorrere una congrua quantità di tempo nello studio al di fuori delle lezioni. Una preparazione soddisfacente richiede in media 1 ora di studio per ciascuna ora trascorsa in aula e circa 80 ore per lo sviluppo del progetto. Il materiale delle lezioni sarà disponibile sulla piattaforma e-learning dipartimentale http://elearning.informatica.unisa.it/el-platform/. Contatti Prof. Giuseppe Polese gpolese@unisa.it