

Principali informazioni sull'insegnamento	
Titolo insegnamento	Data Mining
Corso di studio	Laurea Magistrale in Data Science
Crediti formativi	9 (7 Lezione + 2 Esercitazione)
Denominazione inglese	Data Mining
Obbligo di frequenza	No
Lingua di erogazione	Italiano

Docente responsabile	Nome Cognome	Indirizzo Mail
	Donato Malerba	donato.malerba@uniba.it

Dettaglio crediti formativi	Ambito disciplinare	SSD	Crediti
	Tecnologie dell'Informatica	ING-INF/05	9

Modalità di erogazione	
Periodo di erogazione	Il semestre
Anno di corso	Primo
Modalità di erogazione	Lezioni frontali Esercitazioni guidate in laboratorio

Organizzazione della didattica	
Ore totali	225
Ore di corso	86
Ore di studio individuale	139

Calendario	
Inizio attività didattiche	24 febbraio 2020
Fine attività didattiche	29 maggio 2020

Syllabus	
Prerequisiti	<p>FONDAMENTI DI MATEMATICA: nozioni di insiemistica, numeri interi e reali, funzioni reali di variabile reale (valore assoluto, logaritmo, esponenziale, funzioni trigonometriche), limiti e continuità per funzioni di una variabile, derivate di una funzione in una variabile, integrali di funzioni in una variabile, cenni di combinatorica, calcolo matriciale, autovalori e autovettori, relazioni, relazioni funzionali, di equivalenza e di ordine. Reticolo.</p> <p>FONDAMENTI DI PROGRAMMAZIONE: algoritmi, linguaggi di programmazione e programmi, flusso di controllo, funzioni, strutture dati, stringhe e file, ricorsione, ricerca, ordinamento, complessità computazionale. Python.</p>

	<p>GESTIONE DI DATI STRUTTURATI E NON STRUTTURATI: Dati, Informazioni, Conoscenza. Dati strutturati e non strutturati. Base di dati e Sistema Informatico. Transazioni. Modello Concettuale e sua rappresentazione in diagrammi E/R. Modello logico relazionale. Fondamenti di SQL. Rappresentazione e interrogazione di dati spaziali, temporali e testuali.</p> <p>MODELLIZZAZIONE STATISTICA: distribuzione di probabilità, principali distribuzioni univariate. Probabilità condizionata.</p>
<p>Risultati di apprendimento previsti (declinare rispetto ai Descrittori di Dublino) (si raccomanda che siano coerenti con i risultati di apprendimento del CdS, compreso i risultati di apprendimento trasversali)</p>	<ul style="list-style-type: none"> • <i>Conoscenza e capacità di comprensione</i> <ul style="list-style-type: none"> - Acquisizione di conoscenze relative agli algoritmi di data mining più noti in letteratura. - Comprensione delle scelte di algoritmi di data mining per specifici compiti. - Capacità di interpretazione dei risultati di un algoritmo di data mining. • <i>Conoscenza e capacità di comprensione applicate</i> <ul style="list-style-type: none"> - Capacità di realizzazione di un semplice progetto di scoperta di conoscenza in una collezione di dati mediante: <ul style="list-style-type: none"> - Utilizzo e/o sviluppo di strumenti per la selezione, pre-elaborazione e trasformazione dei dati, e per la validazione dei pattern estratti. - Utilizzo di strumenti di data mining per l'estrazione di conoscenza finalizzata a scopi descrittivi in diversi contesti applicativi (aziendali e scientifici). • <i>Autonomia di giudizio</i> <ul style="list-style-type: none"> - Gli studenti sono in grado di apprezzare l'uso di algoritmi di data mining in processi di scoperta della conoscenza. - L'autonomia di giudizio viene acquisita attraverso lo studio e l'interpretazione critica dei testi. - Il raggiungimento dell'adeguata autonomia è verificato attraverso le esercitazioni, che si tengono durante il corso, e con l'esame finale di profitto. • <i>Abilità comunicative</i> <ul style="list-style-type: none"> - Gli studenti sono in grado di esporre le tematiche incluse nel programma del corso mediante il lessico specifico della disciplina. • <i>Capacità di apprendere</i> <ul style="list-style-type: none"> - Gli studenti sono in grado di approfondire in autonomia le tematiche incluse nel programma del corso anche ricorrendo a risorse non direttamente coinvolte nella erogazione delle ore di lezione.
<p>Contenuti di insegnamento</p>	<p>I. Scoperta di conoscenza nelle basi di dati: il processo. La scoperta di conoscenza nelle basi di dati: definizione. Il processo della scoperta di conoscenza nelle basi di dati. Il processo CRISP-DM: business understanding, data</p>

	<p>understanding, data preparation, modelling, evaluation, deployment.</p> <p>2. Similarità e distanze. Similarità/distanze per dati qualitativi e quantitativi, misure di similarità di testi, misure di similarità temporale.</p> <p>3. Analisi di associazione. insieme frequente, regola di associazione, principali esempi di pattern di associazione. Applicazioni all'analisi di cestino, di log e alla bioinformatica.</p> <p>4. Analisi di raggruppamento (clustering). Algoritmi di clustering partizionale, gerarchico, basati su modelli probabilistici, basati su griglia e densità. Validazione di cluster. Applicazioni a sistemi di raccomandazione, analisi di reti sociali, marketing, biologia e pianificazione urbana.</p> <p>5. Analisi di casi estremi (outlier). Analisi di valori estremi, modelli probabilistici, clustering per l'individuazione di casi estremi, rilevamento di casi estremi basato su distanza e su densità. Validità dei casi estremi. Applicazioni in sicurezza informatica, ambito medico e finanziario.</p> <p>6. Analisi di dati testuali, spaziali e di serie temporali. Algoritmi di clustering per testi e loro applicazioni. Algoritmi di clustering e outlier-detection spaziale e loro applicazioni. Pattern di co-locazione spaziale e raggruppamento di traiettorie. Motivi di serie temporali. Raggruppamenti di serie temporali. Analisi di casi estremi in serie temporali. Applicazioni a organizzazione di testi, topic modeling, analisi di dati provenienti da sensori.</p>
--	---

Programma	
Testi di riferimento	<p>Charu C. Aggarwal Data Mining Springer 2015</p> <p>Annalisa Appice, Anna Ciampi, Fabio Fumarola, Donato Malerba Data Mining Techniques in Sensor Networks: Summarization, Interpolation and Surveillance Springer 2014</p> <p>Jake VanderPlas Python Data Science Handbook. Essential Tools for Working with Data O'Reilly, 2017</p>
Note ai testi di riferimento	I testi di riferimento saranno supportati da articoli scientifici e dispense forniti dal docente durante lo svolgimento del corso
Metodi didattici	<ul style="list-style-type: none"> - Lezioni frontali condotte con l'ausilio di supporti didattici (slide) - Esercitazioni guidate in laboratorio
Metodi di valutazione (indicare almeno la	- Prova scritta sulla parte teorica

tipologia scritto, orale, altro)	<ul style="list-style-type: none"> - Svolgimento di un progetto di scoperta di conoscenza nei dati mediante l'applicazione di algoritmi di data mining. - La prova scritta è propedeutica alla presentazione del progetto.
<p>Criteri di valutazione (per ogni risultato di apprendimento atteso su indicato, descrivere cosa ci si aspetta lo studente conosca o sia in grado di fare e a quale livello al fine di dimostrare che un risultato di apprendimento è stato raggiunto e a quale livello)</p>	<p>Nella prova scritta, lo studente dev'essere in grado di esporre, in modo critico, i concetti appresi relativi al processo di scoperta della conoscenza. Dev'essere altresì capace di affrontare semplici esercizi di data mining.</p> <p>Nel progetto lo studente deve dimostrare capacità di analisi dei dati, di applicazione di algoritmi di data mining e di comprensione dei risultati ottenuti, in un ciclo finalizzato al miglioramento delle prestazioni. Gli strumenti che si possono all'uopo utilizzare sono quelli illustrati durante le ore di esercitazione/laboratorio.</p>
Altro	