



Principali informazioni sull'insegnamento

Denominazione dell'insegnamento	Data Mining	
Corso di studio	Laurea Magistrale in Data Science	
Anno Accademico	2025/26	
Crediti formativi universitari (CFU) / European Credit Transfer and Accumulation System (ECTS)	6 CFU	
Settore Scientifico Disciplinare	IINF-05/A - Sistemi di elaborazione delle informazioni	
Lingua di erogazione	Italiano	
Anno di corso	Primo	
Periodo di erogazione	2^ semestre, le date esatte sono riportate nel manifesto/regolamento	
Obbligo di frequenza	No, ma la frequenza è fortemente raccomandata	
Sito web del corso di studio	https://www.uniba.it/it/ricerca/dipartimenti/informatica/didattica/corsi-di-laurea/corsi-di-laurea	

Docente/i	
Nome e cognome	Donato Malerba
Indirizzo mail	donato.malerba@uniba.it
Telefono	080 5443269
Sede	Dipartimento di Informatica, Via Orabona 4, 70125, Bari. Stanza n.508, 5^ piano.
Sede virtuale	Piattaforma e-learning UNIBA - https://elearning.uniba.it/
Sito web del docente	https://www.uniba.it/it/docenti/malerba-donato
Ricevimento (giorni, orari e modalità, es. su appuntamento)	Mercoledì 11.00-13.00 o su appuntamento

Syllabus	
Obiettivi formativi	Acquisizione di adeguate conoscenze dei processi di scoperta della conoscenza nei



	dati (KDD). Acquisizione di competenza nell'analisi di un dataset con strumenti di data mining.
Prerequisiti	Le seguenti conoscenze preliminari facilitano ed accelerano la comprensione degli argomenti dell'insegnamento: Da 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. Da FONDAMENTI DI PROGRAMMAZIONE: algoritmi, linguaggi di programmazione e programmi, flusso di controllo, funzioni, strutture dati, stringhe e file, ricorsione, ricerca, ordinamento, complessità computazionale. Python. Da 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. Da MODELLIZZAZIONE STATISTICA: distribuzione di probabilità, principali distribuzioni univariate. Probabilità condizionata.
Contenuti di insegnamento (Programma)	1. Scoperta di conoscenza nelle basi di dati: il processo. (42 ore con laboratorio) 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 understanding, data preparation, modelling, evaluation, deployment. 2. Similarità e distanze. (20 ore con laboratorio) Similarità/distanze per dati qualitativi e quantitativi, misure di similarità di testi, misure di similarità temporale.
Testi di riferimento	Charu C. Aggarwal <i>Data Mining</i> Springer 2015 (disponibile in biblioteca e su piattaforma Ada) Gli studenti che lo desiderano possono ottenere i testi in prestito dalla Biblioteca. Può convenire verificarne la disponibilità mediante il Sistema Bibliotecario di Ateneo https://opac.uniba.it/easyweb/w8018/index.php? e contattare la biblioteca per concordare il prestito.
Note ai testi di riferimento	Il testo di riferimento è supportato da articoli scientifici e dispense forniti dal docente durante lo svolgimento del corso e disponibili sulla piattaforma e-learning di UNIBA. Sulla stessa piattaforma sono resi disponibili esempi di analisi di dataset, richiesta per la discussione orale, e tracce di prove scritte/esonero.
Organizzazione della didattica	
Ore	



Totali	Didattica frontale	Laboratorio/esercitazioni	Progetto	Studio individuale
150 ore	32 ore	30 ore	0 ore	88 ore
CFU/ETCS				
6 CFU	4 CFU	2 CFU	0 CFU	

Metodi didattici	
	Le 62 ore previste per l'insegnamento sono ripartite come segue: 32 ore di didattica frontale (in presenza); 30 ore di laboratorio (in presenza).

Risultati di apprendimento previsti	
Conoscenza e capacità di comprensione	● Conoscenze teoriche ○ Comprendere i principi fondamentali della scoperta di conoscenza nelle basi di dati (KDD). ○ Conoscere le fasi del processo CRISP-DM e il loro ruolo nell'analisi dei dati (business understanding, data understanding, data preparation, modelling, evaluation, deployment). ○ Conoscere le principali misure di similarità e distanza per dati quantitativi, qualitativi, testuali e temporali. ○ Comprendere le implicazioni metodologiche e applicative delle diverse metriche di similarità in contesti reali. ● Capacità di comprensione ○ Interpretare e valutare dataset complessi, riconoscendo le problematiche legate alla qualità dei dati e alla loro preparazione. ○ Analizzare criticamente i risultati di un processo di data mining, valutandone la validità e l'applicabilità. ○ Selezionare e giustificare l'uso di misure di similarità/distanza adeguate in funzione della tipologia di dati e del problema da risolvere. ○ Collegare concetti teorici a casi pratici, comprendendo come le tecniche di KDD e le metriche di similarità supportino decisioni basate sui dati.
Conoscenza e capacità di comprensione applicate	● Progettazione e realizzazione di un processo di KDD: ○ Capacità di impostare e condurre un semplice progetto di scoperta di conoscenza su una collezione di dati. ○ Utilizzo e/o sviluppo di strumenti per la selezione, pre-elaborazione e trasformazione dei dati, con attenzione alla qualità e alla coerenza delle informazioni. ○ Validazione dei pattern estratti mediante metriche e tecniche di valutazione appropriate. ● Applicazione di strumenti di Data Mining: ○ Uso di algoritmi e software di data mining per l'estrazione di conoscenza.



	○ Capacità di interpretare i risultati ottenuti e di contestualizzarli in diversi ambiti applicativi, sia aziendali (supporto alle decisioni, analisi di mercato) sia scientifici (analisi di dati sperimentali, bioinformatici, sociali). ● Gestione delle misure di similarità e distanza: ○ Applicazione pratica di metriche di similarità/distanza su dati quantitativi, qualitativi, testuali e temporali. ○ Capacità di scegliere la misura più adeguata in funzione del problema e di giustificarne l'utilizzo. ● Competenze trasversali: ○ Capacità di integrare conoscenze teoriche e pratiche per affrontare problemi reali di analisi dei dati. ○ Sviluppo di un approccio critico e consapevole all'uso degli strumenti, con attenzione alle implicazioni metodologiche e applicative.
Competenze trasversali	Autonomia di giudizio ○ 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. Abilità comunicative ○ Gli studenti sono in grado di esporre le tematiche incluse nel programma del corso mediante il lessico specifico della disciplina, e sono in grado di comunicare in modo chiaro e privo di ambiguità le loro conclusioni nell'analisi di un dataset, a interlocutori specialisti e non specialisti. Capacità di apprendere in modo autonomo ○ Gli studenti sono in grado di approfondire in autonomia le tematiche incluse nel programma del corso anche ricorrendo a risorse (articoli su algoritmi di data mining e/o strumenti di analisi) non direttamente utilizzate nelle ore di lezione/laboratorio.

Valutazione	
Modalità di verifica dell'apprendimento	○ Prova scritta sulla parte teorica. ○ Discussione dell'analisi di un dataset assegnato dal docente, mediante l'applicazione della metodologia CRISP-DM vista nel corso e di strumenti scelti dallo studente, al fine di scoprire pattern e cluster di interesse. La prova scritta è propedeutica alla discussione dell'analisi del dataset. La votazione finale è in trentesimi. Ad essa contribuiscono il voto della prova scritta e la valutazione della discussione dell'analisi del dataset. Sono previste due prove parziali, una a metà corso e una a fine corso, che nell'insieme sostituiscono la prova scritta sulla parte teorica. Le prove parziali sono mantenute per il solo anno accademico in cui sono state sostenute. Durante la prova scritta è possibile utilizzare una calcolatrice, mentre nella discussione dell'analisi del dataset è possibile utilizzare il proprio PC dove sono installati gli ambienti utilizzati per l'analisi del dataset.
Criteri di valutazione	Conoscenza e capacità di comprensione: ○ esposizione critica dei concetti appresi relativi al processo di scoperta della conoscenza e capacità di affrontare semplici esercizi di data mining.



	Conoscenza e capacità di comprensione applicate: ○ analisi di dataset con applicazione di algoritmi di data mining e comprensione dei risultati ottenuti, in un ciclo finalizzato al miglioramento delle prestazioni Autonomia di giudizio: ○ Capacità di svolgere semplici esercizi assegnati durante il corso delle lezioni Abilità comunicative: ○ Uso del lessico specifico della disciplina informatica Capacità di apprendere: ○ sviluppo di argomenti su data mining non direttamente trattati nel corso ma assegnati dal docente
Criteri di misurazione dell'apprendimento e di attribuzione del voto finale	Il voto finale sarà attribuito sulla base della valutazione comparativa di tutti i criteri di valutazione sopra citati
Altro	Si suggerisce agli studenti di affidarsi esclusivamente alle informazioni/comunicazioni fornite sui siti ufficiali del Dipartimento di Informatica, ovvero sui gruppi social solo se costituiti e amministrati esclusivamente dai docenti dei relativi insegnamenti: ● https://www.uniba.it/it/ricerca/dipartimenti/informatica/didattica/corsi-di-laurea ● https://www.uniba.it/it/ricerca/dipartimenti/informatica ● https://elearning.uniba.it/ I programmi di tutti gli insegnamenti sono disponibili al seguente link: ● https://elearning.uniba.it/course/index.php?categoryid=2277 Le informazioni che tutti gli studenti dovrebbero conoscere sono scritte nei regolamenti didattici dei Corsi di Studi disponibili nel sito: ● https://www.uniba.it/it/ricerca/dipartimenti/informatica/didattica/corsi-di-laurea Si suggerisce agli studenti di diffidare delle informazioni e dei materiali circolanti su siti o gruppi social non ufficiali, poiché spesso sono risultati non affidabili, non corretti o incompleti. Per ogni dubbio, chiedere un incontro al docente secondo le modalità previste per il ricevimento. Link al corso sulla piattaforma e-learning ADA: ● https://elearning.uniba.it/course/index.php?categoryid=2277



Main information on the course

Course name	Data Mining	
Degree	Master's Degree in Data Science	
Academic year	2025/26	
European Credit Transfer and Accumulation System (ECTS), in Italian Crediti Formativi Universitari (CFU)	6 CFU (each CFU corresponds to 25 hours (h) of student's time); CFU are of type T1, T2 or T3 T1 = 8 h lecture + 17 h individual study T2 = 15 h practice + 10 h individual study T3 = 25 h individual study	
Settore Scientifico Disciplinare	IINF-05/A - Sistemi di elaborazione delle informazioni	
Course language	Italian	
Course year	First	
Course period	Second Semester - exact dates can be found in the didactic regulations	
Course attendance requirement	None, but it is highly recommended to attend classes	
Website of the Degree	https://www.uniba.it/it/ricerca/dipartimenti/informatica/didattica/corsi-di-laurea/corsi-di-laurea	

Teacher(s)

Name and Surname	Donato Malerba
email	donato.malerba@uniba.it
phone	080 5443269
office	Dipartimento di Informatica, Via Orabona 4, 70125, Bari. Room n.508, 5th floor
e-learning platform	Platform ADA - https://elearning.uniba.it/
Teacher's homepage	https://www.uniba.it/it/docenti/malerba-donato
Office hours	Wednesday 11:00-13:00 or on appointment

Syllabus

Course goals	Acquisition of adequate knowledge of Knowledge Discovery in Databases (KDD) processes. Gaining proficiency in analyzing a dataset using data mining tools, specifically in the utilization of clustering algorithms and association rule discovery.
Prerequisites/requirements	The following preliminary knowledge facilitates and accelerates the understanding of the course topics: From FUNDAMENTALS OF MATHEMATICS: Notions of set theory, integers and real numbers, real functions of a real variable (absolute value, logarithm, exponential, trigonometric functions), limits and continuity for functions of one variable, derivatives of a function in one variable, integrals of functions in one variable, basics of combinatorics, matrix calculus, eigenvalues and eigenvectors, relations, functional relations, equivalence relations, and order relations. Lattice. From FUNDAMENTALS OF PROGRAMMING: Algorithms, programming languages, and programs, flow control, functions, data structures, strings, and files, recursion, search, sorting, computational complexity. Python. From MANAGEMENT OF STRUCTURED AND UNSTRUCTURED DATA:



	Data, Information, Knowledge. Structured and unstructured data. Databases and Information Systems. Transactions. Conceptual model and its representation in E/R diagrams. Logical relational model. Basics of SQL. Representation and querying of spatial, temporal, and textual data.			
	From STATISTICAL MODELING: Probability distribution, main univariate distributions. Conditional probability.			
Course program	1. Knowledge Discovery in Databases (KDD): The Process. (42 hours including practice sessions) Knowledge Discovery in Databases: Definition. The process of Knowledge Discovery in Databases. The CRISP-DM process: business understanding, data understanding, data preparation, modeling, evaluation, deployment. 2. Similarity and Distances. (20 hours including practice sessions) Similarity/distances for qualitative and quantitative data, measures of text similarity, measures of temporal similarity.			
Books of reference	Charu C. Aggarwal <i>Data Mining</i> Springer 2015 (available in the library and on the Ada platform) Students who wish to can borrow the texts from the Library. It may be advisable to check their availability through the University Library System at https://opac.uniba.it/easyweb/w8018/index.php? and contact the library to arrange the loan.			
Notes to the books	The reference texts are supported by scientific articles and slides provided by the instructor during the course.			
Organization of the didactic activities				
Hours				
Total	Lectures	Practice sessions	Project work	Individual study
150 hours	32 hours	30 hours	0 hours	88 hours
CFU/ETCS				
6 CFU	4 CFU	2 CFU	0 CFU	
Teaching methods				
	The 62 hours allocated for classes are distributed as follows: - 32 hours of lectures (in person); - 30 hours of practice sessions (in person).			
Expected learning outcomes				
Knowledge and understanding	• Theoretical Knowledge ○ Understand the fundamental principles of Knowledge Discovery in Databases (KDD). ○ Know the phases of the CRISP-DM process and their role in data analysis (business understanding, data understanding, data preparation, modelling, evaluation, deployment). ○ Know the main similarity and distance measures for quantitative,			



	○ qualitative, textual, and temporal data. ○ Understand the methodological and applicative implications of different similarity metrics in real-world contexts. ● Comprehension Skills ○ Interpret and evaluate complex datasets, recognizing issues related to data quality and preparation. ○ Critically analyze the results of a data mining process, assessing their validity and applicability. ○ Select and justify the use of appropriate similarity/distance measures depending on the type of data and the problem to be solved. ○ Connect theoretical concepts to practical cases, understanding how KDD techniques and similarity metrics support data-driven decision making.
Applying knowledge and understanding	● Design and implementation of a KDD process ○ Ability to set up and carry out a simple knowledge discovery project on a data collection. ○ Use and/or development of tools for data selection, preprocessing, and transformation, with attention to data quality and consistency. ○ Validation of extracted patterns through appropriate metrics and evaluation techniques. ● Application of Data Mining tools ○ Use of data mining algorithms and software for knowledge extraction. ○ Ability to interpret the obtained results and contextualize them in different application domains, both business (decision support, market analysis) and scientific (analysis of experimental, bioinformatics, and social data). ● Management of similarity and distance measures ○ Practical application of similarity/distance metrics on quantitative, qualitative, textual, and temporal data. ○ Ability to select the most appropriate measure depending on the problem and justify its use. ● Cross-cutting skills ○ Ability to integrate theoretical and practical knowledge to address real-world data analysis problems. ○ Development of a critical and informed approach to the use of tools, with attention to methodological and applicative implications.
Other skills	<i>Making judgements</i> ○ Students are able to appreciate the use of data mining algorithms in knowledge discovery processes. ○ Independent judgment is acquired through the study and critical interpretation of texts. ○ The achievement of adequate autonomy is assessed through exercises held during the course and the final exam. <i>Communication</i> ○ Students are able to present the topics included in the course syllabus using the specific terminology of the discipline, and they are capable of clearly and unambiguously communicating their conclusions in the analysis of a dataset to both specialist and non-specialist audiences. <i>Learning skills</i> ○ Students are able to independently delve deeper into the topics covered in the course syllabus, even by using resources (such as articles on data mining algorithms and/or analysis tools) that are not directly explained during lectures or practice sessions.



Assessment	
Assessment methods	○ Written Exam on Theoretical Content: A written exam covering the theoretical part of the course. ○ Discussion of a Dataset Analysis: Students must discuss the analysis of a dataset assigned by the instructor, applying the CRISP-DM methodology covered during the course and using tools chosen by the student. The goal is to discover patterns and clusters of interest. The written exam is a prerequisite for the discussion of dataset analysis. The final grade is a vote between 0 and 30 (cum laude) and is based on the score of the written exam and the evaluation of the dataset analysis discussion. Two partial exams are offered, one at mid-course and one at the end of the course, which together replace the written exam on the theoretical content. These partial exams are valid only for the academic year in which they are taken. During the written exam, students are allowed to use a pocket calculator, while during the dataset analysis discussion, they can use their own PC with the necessary software installed for dataset analysis.
Evaluation criteria	Knowledge and Understanding: ○ Critical presentation of the concepts learned related to the knowledge discovery process and the ability to tackle simple data mining exercises. Applied Knowledge and Understanding: ○ Analysis of datasets with the application of data mining algorithms and understanding of the obtained results, in a cycle aimed at improving performance. Autonomy of Judgment: ○ Ability to carry out simple exercises assigned during the course lectures. Communication Skills: ○ Use of specific vocabulary in the field of computer science. Ability to Learn: ○ Development of topics on data mining not directly covered in the course but assigned by the instructor.
Measurements and final grade	The final grade will be awarded based on the comparative evaluation of all the assessment criteria mentioned above.
Further information	It is recommended that students rely exclusively on information and communications provided on the official websites of the Department of Computer Science or on social groups if they are established and managed solely by the instructors of the respective courses: ● https://www.uniba.it/it/ricerca/dipartimenti/informatica/didattica/corsi-di-laurea/corsi-di-laurea ● https://www.uniba.it/it/ricerca/dipartimenti/informatica



- <https://elearning.uniba.it/>

The course syllabi are available here:

- <https://elearning.uniba.it/course/index.php?categoryid=2277>

The information that all students should know is detailed in the educational regulations and course handbooks available on the website:

- <https://www.uniba.it/it/ricerca/dipartimenti/informatica/didattica/corsi-di-laurea/corsi-di-laurea>

Students are advised to be cautious regarding information and materials circulating on unofficial websites or social groups, as they can often be unreliable, incorrect, or incomplete. For any doubts, students should arrange a meeting with the instructor following the specified office hours.

Link to the course on the ADA e-learning platform:

- <https://elearning.uniba.it/course/index.php?categoryid=2277>