Principali informazioni sull'insegnamento

Denominazione dell'insegnamento	**Data Mining**
Corso di studio	Laurea Magistrale in Data Science
Anno Accademico	2024/25
Crediti formativi universitari (CFU) / European Credit Transfer and Accumulation System (ECTS)	9 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) e di fondamentali tecniche di data mining per analisi di raggruppamento

	e di associazione. Acquisizione di competenza nell'analisi di un dataset con strumenti di data mining, specificatamente nell'utilizzo di algoritmi di clustering e scoperta di regole di associazione.
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. (36 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. (18 ore con laboratorio) Similarità/distanze per dati qualitativi e quantitativi, misure di similarità di testi, misure di similarità temporale. 3. Analisi di associazione. (18 ore con laboratorio) insieme frequente, regola di associazione, principali esempi di pattern di associazione. Applicazioni all'analisi di cestino, di log e alla bioinformatica. 4. Analisi di raggruppamento (clustering). (14 ore con laboratorio) 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.
Testi di riferimento	Charu C. Aggarwal *Data Mining* 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	I testi di riferimento sono supportati 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/esercitazione	Progetto	Studio individuale
225 ore	56 ore	30 ore	0 ore	139 ore

CFU/ETCS				
9 CFU	7 CFU	2 CFU	0 CFU	

Metodi didattici

Le 86 ore previste per l'insegnamento sono ripartite come segue:
56 ore di didattica frontale (in presenza);
30 ore di laboratorio (in presenza).

Risultati di apprendimento previsti

Conoscenza e capacità di comprensione

- 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.

Conoscenza e capacità di comprensione applicate

- Capacità di realizzazione di un semplice progetto di scoperta di conoscenza in una collezione di dati mediante:
  - 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).

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 acquista 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=456

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=456

Main information on the course

Course name	**Data Mining**	
Degree	Master's Degree in Data Science	
Academic year	2024/25	
European Credit Transfer and Accumulation System (ECTS), in Italian Crediti Formativi Universitari (CFU)		9 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		
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 and fundamental data mining techniques for cluster analysis and association analysis. 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. (36 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. (18 hours including practice sessions) Similarity/distances for qualitative and quantitative data, measures of text similarity, measures of temporal similarity. 3. Association Analysis. (18 hours including practice sessions) Frequent itemsets, association rules, main examples of association patterns. Applications in market basket analysis, log analysis, and bioinformatics. 4. Clustering Analysis. (14 hours including practice sessions) Partitioning, hierarchical, probabilistic model-based, grid-based, and density-based clustering algorithms. Cluster validation. Applications in recommendation systems, social network analysis, marketing, biology, and urban planning.
Books of reference	Charu C. Aggarwal *Data Mining* 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
225 hours	56 hours	30 hours	0 hours	139 hours
CFU/ETCS				
9 CFU	7 CFU	2 CFU	0 CFU	

Teaching methods

The 86 hours allocated for classes are distributed as follows:

- 56 hours of lectures (in person);
- 30 hours of practice sessions (in person).

Expected learning outcomes

Knowledge and understanding	○ Acquisition of knowledge related to the most well-known data mining algorithms in the literature. ○ Understanding the choices of data mining algorithms for specific tasks. ○ Ability to interpret the results of a data mining algorithm.
Applying knowledge and understanding	○ Ability to carry out a simple knowledge discovery project in a data collection through: ▪ Utilization and/or development of tools for data selection, pre-processing, transformation, and validation of extracted patterns. ▪ Usage of data mining tools for knowledge extraction aimed at descriptive purposes in various application contexts (business and scientific).
Other skills	*Making judgements* ○ 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. *Communication* ○ 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. *Learning skills* ○ 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 the 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 environments 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 only 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=456 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: hhttps://elearning.uniba.it/course/index.php?categoryid=456