UNIVERSITÀ DEGLI STUDI DI PALERMO

DIPARTIMENTO	Scienze Economiche, Aziendali e Statistiche		
ANNO ACCADEMICO OFFERTA	2024/2025		
ANNO ACCADEMICO EROGAZIONE	2024/2025		
CORSO DILAUREA MAGISTRALE	STATISTICA E DATA SCIENCE		
INSEGNAMENTO	BIG DATA C.I.		
CODICE INSEGNAMENTO	22216		
MODULI	Si		
NUMERO DI MODULI	2		
SETTORI SCIENTIFICO-DISCIPLINARI	ING-INF/05		
DOCENTE RESPONSABILE	PIRRONE ROBERTO	Professore Ordinario	Univ. di PALERMO
ALTRI DOCENTI	PIRRONE ROBERTO	Professore Ordinario	Univ. di PALERMO
	LA CASCIA MARCO	Professore Ordinario	Univ. di PALERMO
CFU	12		
PROPEDEUTICITA'			
MUTUAZIONI			
ANNO DI CORSO	1		
PERIODO DELLE LEZIONI	Annuale		
MODALITA' DI FREQUENZA	Facoltativa		
TIPO DI VALUTAZIONE	Voto in trentesimi		
ORARIO DI RICEVIMENTO DEGLI STUDENTI	**LA CASCIA MARCO** Lunedì 15:00 17:00 Microsoft Teams Codice: wztkv0u **PIRRONE ROBERTO** Mercoledì 11:30 13:00 Studio del docente, Edificio 6, terzo piano, stanza 3025		

DOCENTE: Prof. ROBERTO PIRRONE

PREREQUISITI	Concetti base di Statistica
RISULTATI DI APPRENDIMENTO ATTESI	Conoscenza e capacita' di comprensione Lo studente, al termine del corso, avra' acquisito conoscenze e metodologie per affrontare le problematiche legate sia all'analisi dei piu' diffusi tipi di dati sia all'utilizzo di architetture software per la gestione dei Big Data. Lo studente conoscera' in maniera adeguata le differenze tra i diversi algoritmi di analisi in relazione alla tipologia dei dati, conoscera' le tecniche di pre-processing piu' adatte e come definire l'architettura Big Data piu' efficiente per condurre le proprie analisi. Per il raggiungimento di quest'obiettivo il corso comprende un ciclo di lezioni frontali sugli argomenti della disciplina. Per la verifica di quest'obiettivo l'esame comprende una serie di domande teoriche nella prova scritta relativa a ciascun modulo e la discussione orale dei risultati della stessa. Capacita' di applicare conoscenza e comprensione Lo studente avra' acquisito conoscenze e metodologie per analizzare e risolvere problemi tipici legati alla l'implementazione di pipeline complete di analisi dei dati sia per dataset classici sia per Big Data. Egli avra' profonda conoscenza del linguaggio di programmazione Python e delle principali librerie per l'analisi e la visualizzazione dei dati quali Numpy, SciPy, Scikit-learn, Matplotlib, Pandas, Tensorflow e Keras. Inoltre lo studente avra' sufficiente conoscenza dei database noSQL quale Apache Cassandra e del framework per Big Data Apache Hadoop con il suo ecosistema, mentre acquisira' profonda conoscenza del farmework Apache Spark e delle sue librerie di analisi dei dati e di interazione con i database nella loro interfaccia Python. Per il raggiungimento di quest'obiettivo il corso comprende una serie di esercitazioni teoriche per sviluppo di pipeline di analisi dei dati. Per la verifica di quest'obiettivo l'esame comprende una serie di domande applicative nella prova scritta relativa a ciascun modulo e la discussione orale dei risultati della stessa. Autonomia di giudizio Lo studente sara' in grado di svolgere un'analisi comparativa delle caratteristiche di differenti ambienti e/o infrastrutture di analisi di Big Data in relazione alla soluzione di problemi specifici. Egli sara' in grado di affrontare a livello operativo problemi non strutturati e prendere decisioni in regime d'incertezza. Attraverso l'approccio metodologico acquisito durante il corso, egli potra' condurre lo sviluppo di nuove problematiche applicative nell'ambito dei Big Data e della data analysis in generale. Per il raggiungimento di quest'obiettivo il corso comprende una serie di esercitazioni teoriche in ciascun modulo. Per la verifica di quest'obiettivo l'esame comprende una serie di domande teoriche nella prova scritta relativa a ciascun modulo e la discussione orale dei risultati della stessa. Abilita' comunicative Lo studente sara' in grado di comunicare con competenza e proprieta' di linguaggio problematiche complesse di data analysis e Big Data. Per il raggiungimento di quest'obiettivo il corso comprende una serie di esercitazioni teoriche in ciascun modulo. Per la verifica di quest'obiettivo l'esame comprende la discussione orale dei risultati della prova scritta relativa a ciascun modulo. Capacita' d'apprendimento Lo studente sara' in grado di affrontare in autonomia qualsiasi problematica concernente lo sviluppo di pipeline complete per analisi di Big Data. Sara' in grado di approfondire tematiche complesse legate all'analisi di prestazioni di framework diversi cogliendone i punti di forza e di debolezza. Per il raggiungimento di quest'obiettivo il corso comprende una serie di esercitazioni teoriche in ciascun modulo. Per la verifica di quest'obiettivo l'esame comprende la discussione orale dei risultati della prova scritta relativa a ciascun modulo.
VALUTAZIONE DELL'APPRENDIMENTO	L'esame finale consta di due prove scritte separate, una per ciascun modulo, ciascuna seguita da un colloquio orale per la discussione del risultato della relativa prova scritta. Ogni prova avrà la durata di due ore e sarà intesa a valutare sia il grado di conoscenza teorica degli argomenti coperti dal corso sia la competenza pratica raggiunta nell'affrontare i temi coperti dalle esercitazioni. Le domande di tipo teorico saranno aperte mentre i quesiti di tipo applicativo richiederanno la scrittura di codice atto a risolverli. Non si potrà accedere al colloquio orale se lo studente non avrà raggiunto una valutazione di almeno 18/30 nella prova scritta. L'articolazione del voto di esame sarà strutturata per fasce di valutazione:

	-18/30 – 20/30: lo studente ha una conoscenza appena sufficiente dei contenuti teorici dell'insegnamento ed è in grado di sviluppare solo alcune parti degli esercizi di tipo applicativo -21/30 – 23/30: lo studente ha una discreta conoscenza dei contenuti teorici dell'insegnamento; egli riesce a sviluppare sommariamente tutte le componenti richieste dagli esercizi di tipo applicativo -24/30 – 26/30: lo studente ha buona conoscenza dei contenuti teorici dell'insegnamento e sviluppa interamente tutte le componenti richieste dagli esercizi di tipo applicativo -27/30 – 30/30: lo studente ha piena conoscenza dell'insegnamento e sviluppa completamente e correttamente tutte le componenti richieste dagli esercizi di tipo applicativo -30 e lode: lo studente conosce ottimamente gli argomenti teorici del corso e ha ottime capacità di sviluppo di tutte le componenti richieste dagli esercizi di tipo applicativo; egli inoltre mostra originalità e capacità di approfondimento autonomo dei temi trattati: le sue soluzioni di tipo applicativo sono altresì originali.
ORGANIZZAZIONE DELLA DIDATTICA	Lezioni frontali; Esercitazioni teoriche; Esercitazioni di gruppo per lo sviluppo di pipeline di analisi dei dati con tecnologie Big Data.

MODULO
TECNOLOGIE PER I BIG DATA

Prof. MARCO LA CASCIA

TESTI CONSIGLIATI

Data Mining: The Textbook, 2015, Charu C. Aggarwal, Springer-Verlag New York, ISBN 978-3319141411, disponibile gratuitamente in forma elettronica per gli studenti dell'Ateneo.
Spark: The Definitive Guide: Big Data Processing Made Simple, 2018, di Bill Chambers e Matei Zaharia, Oreilly & Associates Inc, ISBN 978-1491912218, prezzo orientativo € 45,00.
Introduzione a Python. Per l'informatica e la data science, 2021, di Paul Deitel & Harvey Deitel, Pearson, ISBN 978-8891915924, prezzo orientativo € 45,00.
Note fornite dal docente/Lecture notes

TIPO DI ATTIVITA'	B
AMBITO	70297-Formazione informatica e dell'informazione
NUMERO DI ORE RISERVATE ALLO STUDIO PERSONALE	96
NUMERO DI ORE RISERVATE ALLE ATTIVITA' DIDATTICHE ASSISTITE	54

OBIETTIVI FORMATIVI DEL MODULO

Il corso fornisce agli studenti una conoscenza approfondita delle architetture software per i Big Data nonche' dei principali algoritmi di analisi dei dati e delle tecniche di preprocessing di tali dati, al fine di sviluppare autonomamente intere pipeline di analisi per dei casi di studio reali.
Il modulo consente di acquisire 6 CFU e consta di una serie di lezioni ed esercitazioni teoriche.
Il ciclo di lezioni teoriche presenta dapprima un'introduzione al processo di analisi dei dati nel suo complesso.
Successivamente si affrontano le tecniche di prerocessing dei dati quali la riduzione di dimensionalita' e la gestione di dati mancanti e si introducono alcune misure di similarita' piu' diffusamente usate nel campo della data analysis e algoritmi per l'individuazione di pattern ricorrenti. Si passa quindi ad affrontare le architetture software per i Big Data: si affronteranno i database noSQL, il paradigma MapReduce e Apache Hadoop e il framework Apache Spark.
Le esercitazioni prevedono lo studio del linguaggio Python con i moduli numpy, pandas, matplotlib e sklearn, la configurazione degli ambienti di sviluppo con cui si operera' durante il corso e l'implementazione di alcune delle tecniche studiate.

PROGRAMMA

ORE	Lezioni
2	Introduzione al Corso. Il processo di analisi dei dati: raccolta dei dati, pre-processing, applicazione delle tecniche di analisi ed estrazione della conoscenza.
3	Preparazione dei dati: tipi di dati, data cleaning, gestione dei dati mancanti, campionamento.
3	Riduzione della dimensionalita: Principal Component Analysis, Singular Value Decomposition, Trasformazioni Wavelet, Multi Dimensional Scaling, Embedding di grafi.
4	Distanze e similarita' per i diversi tipi di dati: dati quantitativi, dati categoriali, dati testuali, sequenze temporali, grafi.
4	Mining di pattern ricorrenti: algoritmo Apriori, misure statistiche di correlazione.
4	Architetture software per i Big Data: database noSQL, MongoDB. Data lake.
8	Architetture software per i Big Data: l'algoritmo MapReduce, Apache Hadoop, HDFS
8	Architetture software per i Big Data: Spark e le sue librerie.

ORE	Esercitazioni
9	Richiami di Python e dei moduli numpy, pandas, matplotlib, sklearn
3	MongoDB
3	Apache Hadoop, HDFS
3	Analisi dei dati con Spark SQL

MODULO
ANALISI PER BIG DATA

Prof. ROBERTO PIRRONE

TESTI CONSIGLIATI

Data Mining: The Textbook, 2015, Charu C. Aggarwal, Springer-Verlag New York, ISBN 978-3319141411, prezzo orientativo € 70,00

Deep Learning, (2016), di Ian Goodfellow, Yoshua Bengio, Aaron Courville, MIT Press, ISBN 978-0262035613, prezzo orientativo €65,00

Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, 2nd Edition, (2017) Sebastian Raschka, Vahidm Mirjalili, Packt Publishing, ISBN 978-1787125933, prezzo orientativo € 35,00

Spark: The Definitive Guide: Big Data Processing Made Simple, 2018, di Bill Chambers e Matei Zaharia, Oreilly & Associates Inc, ISBN 978-1491912218, prezzo orientativo € 45,00.

Repository per slide ed esercitazioni:
https://github.com/fredffsixty/Big-Data

Siti web con manuali di riferimento per le esercitazioni ed i testi:
https://link.springer.com/book/10.1007%2F978-3-319-14142-8
http://www.deeplearningbook.org/
https://github.com/PacktPublishing/Python-Machine-Learning-Second-Edition
https://github.com/databricks/Spark-The-Definitive-Guide

TIPO DI ATTIVITA'	B
AMBITO	70297-Formazione informatica e dell'informazione
NUMERO DI ORE RISERVATE ALLO STUDIO PERSONALE	96
NUMERO DI ORE RISERVATE ALLE ATTIVITA' DIDATTICHE ASSISTITE	54

OBIETTIVI FORMATIVI DEL MODULO

Il modulo fornisce agli studenti una conoscenza approfondita dei principali algoritmi di analisi dei dati sia nel contesto Big Data sia nel classico contesto del Machine Learning al fine di sviluppare autonomamente intere pipeline di analisi per dei casi di studio reali.

Il modulo consente di acquisire 6 CFU e consta di una serie di lezioni ed esercitazioni teoriche.

Il ciclo di lezioni teoriche presenta dapprima un'introduzione alle Teorie della Probabilità e dell'Informazione nonché ai concetti di stima statistica e campionamento. Si passa poi alla parte del corso dedicata propriamente al machine learning e si affrontano clustering e classificatori nonche' le reti neurali e il deep learning. Infine si presentano alcuni scenari applicativi di interesse quali l'analisi delle immagini mediche, l'elaborazione del linguaggio naturale e l'analisi dei dati web.

Le esercitazioni teoriche coprono l'utilizzo delle librerie Pyhton sci-kit learn, Spark ML e Tensorflow per l'illustrazione dei temi affrontati nel corso teorico attraverso esempi svolti.

PROGRAMMA

ORE	Lezioni
3	Cenni di Teoria della Probabilità e Teoria dell'Informazione; stimatori statistici e tecniche di campionamento.
2	Introduzione al Machine Learning: apprendimento supervisionato, non supervisionato, apprendimento con rinforzo, capacità del modello, parametri e iperparametri, tipologie di errore, tecniche di addestramento.
5	Clustering: k-means e sue varianti, clustering gerarchico, clustering density based e a griglia, clustering basato su grafi, clustering di dati ad elevata dimensionalità, validazione del clustering, analisi degli outlier.
5	Classificatori: feature selection, decision tree e classificatori a regole, Naive Bayes, regressione logistica,Support Vector Machines, Nearest Neighbor, valutazione dei classificatori.
2	Classificatori, concetti avanzati: Multi-class e rare class learning, regressione su dati numerici, semi-supervised learning, metodi di ensemble.
8	Deep Learning: struttura di una rete neurale, tipologia di unità nascoste e di uscita, funzioni di loss, concetto di grafo di computazione, stochastic gradient descent, ottimizzazione e regolarizzazione, CNN, Autoencoder, LSTM, GAN, Graph Neural Networks, fine tuning e transfer learning.
3	Elaborazione di immagini mediche: segmentazione di volumi TAC/RM con CNN 3D.
3	Elaborazione del linguaggio naturale: classificazione di testi con Word2Vec.
5	Analisi di dati web: algoritmo PageRank, recommender systems, web usage analysis, social network analysis.

ORE	Esercitazioni
3	Stima statistica di una distribuzione Gaussiana al variare del tipo di campionamento
3	Implementazione di un algoritmo di clustering in sci-kit learn
3	Implementazione di un algoritmo di classificazione in sci-kit learn
3	Creazione di una pipeline con Spark ML per il clustering
3	Creazione di una pipeline con Spark ML per la classificazione
3	Uso di Tensorflow ed esempi di implementazioni di semplici DNN