

## ADA.14.01.20 - DATA SCIENCE AND ANALYTICS

### SCHEDA DI CASO

**RISULTATO ATTESO 2** - Realizzazione di analisi avanzate di data set (anche di notevoli dimensioni e di big data), anche a partire dalla creazione di modelli, algoritmi e strumenti di raccolta e analisi dati, provvedendo, successivamente, a realizzare rapporti per supportare i processi decisionali

## 1 - PROCEDURE DATA CURATION AND DATA QUALITY CONTROL

Grado di complessità 4

### 1.4 EXPLORATION

Assicurare una maggiore comprensione analitica dei dati raccolti in funzione degli obiettivi delle analisi, in modo da avere una caratterizzazione più accurata delle informazioni disponibili

Grado di complessità 3

### 1.3 CLEANING

Individuare e correggere i dati grezzi (raw data) provenienti da fonti diversi che contengono imprecisioni, armonizzando le relative informazioni mediante specifici dataset (es. Alteryx, Apache Spark, SAS, Big ML, MATLAB, Jupyter)

Grado di complessità 2

### 1.2 ARRICCHIMENTO DEL DATABASE

Categorizzare i dati disponibili a seconda del loro uso finale creando nuove variabili dai valori originali, al fine di consentire un loro facile utilizzo

Grado di complessità 1

### 1.1 STANDARDIZZAZIONE

Promuovere una standardizzazione del processo di data curation, attraverso l'utilizzo di processi semi-automatici (ontologies) di text mining per identificare eventuali sovrapposizioni di metadati

## 2 - CREAZIONE DI MODELLI, ALGORITMI E STRUMENTI PER LA RACCOLTA E ANALISI DEI DATI

Grado di complessità 4

### 2.4 CLASSIFICATION MODELING

Assegnare ogni campione alla giusta categoria, avendo a disposizione una serie di dati appartenenti

## ADA.14.01.20 - DATA SCIENCE AND ANALYTICS

ad un numero finito di categorie note

Grado di complessità 3

### 2.3 CLUSTERING MODELING

Creare i gruppi in modo tale che gli elementi al loro interno abbiano dei punti in comune, avendo a disposizione una serie di dati appartenenti ad un numero finito di gruppi non noti

Grado di complessità 2

### 2.2 NATURAL LANGUAGE PROCESSING

Sfruttare le potenzialità dei modelli di elaborazione del linguaggio naturale (NLP) per il riconoscimento e l'estrazione automatica di informazioni

Grado di complessità 1

### 2.1 FORECASTING MODELING

Esaminare le caratteristiche chiave di una serie ordinata di campioni, per rilevare la stagionalità delle tendenze, l'influenza di fattori esterni, e prevedere i valori futuri di una variabile in base alla sua relazione con le altre o con determinate serie storiche

## 3 - CURA DEI RAPPORTI DI SINTESI E DI COMUNICAZIONE DELL'ESITO DELL'ANALISI A SUPPORTO DELLE DECISIONI

Grado di complessità 5

### 3.5 DATA VISUALISATION

Estrarre e interpretare le informazioni mediante l'utilizzo di strumenti di data visualisation (es. Tableau, RawGraphs, ChartBlocks, PowerBI, QlikSense, ecc.), così da comunicare i risultati con i propri stakeholders in modo efficace

Grado di complessità 4

### 3.4 DESCRIPTIVE ANALYTICS

Predisporre analisi "descrittive" con riferimento a dati riguardanti situazioni attuali o precedenti relativi al mercato, identificando i principali KPI e gli indicatori di prestazione

Grado di complessità 3

### 3.3 PREDICTIVE ANALYTICS

## ADA.14.01.20 - DATA SCIENCE AND ANALYTICS

Predisporre analisi “predittive”, mediante l'utilizzo di tool programmati per fornire risposte relative alla tendenze del mercato nel futuro e avvalendosi di strumenti matematici (forecasting; regressione, applicazione modelli predittivi)

Grado di complessità 2

### 3.2 PRESCRIPTIVE ANALYTICS

Predisporre analisi “prescrittive”, proponendo soluzioni operative/strategiche, sulla base delle analisi svolte, fra le quali il decision maker potrà scegliere

Grado di complessità 1

### 3.1 AUTOMATED ANALYTICS

Predisporre analisi “automatizzate”, capaci di mettere in atto autonomamente le scelte che, in base alle analisi svolte, ritengono valide rispetto agli obiettivi prefissati

## ADA.14.01.20 - DATA SCIENCE AND ANALYTICS

### SCHEDA RISORSE A SUPPORTO DELLA VALUTAZIONE DEL RISULTATO ATTESO 1

#### RISORSE FISICHE ED INFORMATIVE TIPICHE (IN INPUT E/O PROCESS ALLE ATTIVITÀ)

- Obiettivi di analisi
- Norme in materia di protezione della privacy e protezione dati personali (GDPR)
- Linguaggio di programmazione e framework (Python, R)
- Data Analytics (Alteryx, Apache Spark, SAS, Big ML, MATLAB, Jupyter)
- Data Visualisation (Tableau, RawGraphs, ChartBlocks, PowerBI, QlikSense)
- Structured data
- Unstructured data
- Dati grezzi (.pdf, .csv, .txt)

#### TECNICHE TIPICHE DI REALIZZAZIONE/CONDUZIONE DELLE ATTIVITÀ

- Processi semi-automatici (ontologies) di text mining
- Tecniche di data cleaning, exploration
- Natural Language Processing (NLP)
- Forecasting modeling
- Clustering modelling
- Classification modelling
- Tecniche di data visualisation
- Predictive analytics
- Prescriptive analytics
- Automated analytics

#### OUTPUT TIPICI DELLE ATTIVITÀ

- Standardizzazione del processo di data curation svolta
- Data curation e data quality control effettuate
- Modelli, algoritmi e strumenti per l'analisi avanzate di dataset effettuati
- Rapporto finale redatto

#### INDICAZIONI A SUPPORTO DELLA SCELTA DEL METODO VALUTATIVO E DELLA PREDISPOSIZIONE DELLE PROVE

##### ESTENSIONE SUGGERITA DI VARIETÀ PRESTAZIONALE

1. Le diverse tipologie di data curation e data quality control
2. L'insieme degli strumenti per la raccolta e l'analisi dei dati (Natural Language Processing (NLP), Forecasting modeling, Clustering modelling, Classification modelling)
3. L'insieme di tecniche e tools per un'efficace visualizzazione dei dati (data visualisation)
4. Un set di obiettivi di analisi
5. Un set di dati grezzi

## ADA.14.01.20 - DATA SCIENCE AND ANALYTICS

### DISEGNO TIPO DELLA VALUTAZIONE

1. Prova prestazionale: a partire da dataset, rappresentazione del processo di data curation e quality control, creazione di modelli e algoritmi ed illustrazione dei risultati mediante procedure di data visualisation
2. Colloquio tecnico relativo al processo di data curation e quality control, elaborazione e reporting mediante data visualisation per una tipologia di analisi non oggetto di prova prestazionale

## ADA.14.01.20 - DATA SCIENCE AND ANALYTICS

## ADA.14.01.20 - DATA SCIENCE AND ANALYTICS

### FONTI

Cesarini, M., Fontana, M. Mercurio, F., Mezzanecanica M. & Vegetti, N., Data Quality: Un Approccio Metodologico e Applicativo. "Il caso delle COB del mercato del lavoro in Lombardia", Working Paper "Documentazione relativa alla ricostruzione di una metodologia unificata, ripetibile e aperta, CRISP E ARIFL EMC (2015), Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, John Wiley & Sons, Indianapolis

European e-Competence Framework (E-CF), v.04: <https://www.ecompetences.eu/>

Freitas, A. & Curry E. (2016). Big Data Curation, In Cavanillas, J.M & Curry E. (2016). New Horizons for a Data-Driven Economy. A Roadmap for Usage and Exploitation of Big Data in Europe, Big Data Public Private Forum, Springer Open, p. 87-119

Gelernter, J. & Lesk, M. (2011), Use of Ontologies for Data Integration and Curation. In The International Journal of Digital Curation, Issue 1, Volume 6

### SITOGRAFIA

MediaSapens, Data Curation: the essential step for integrated data-driven research