Lezione 7 modulo 1

In questo modulo, parleremo di analisi dei dati principalmente, in particolare di Data Mining e di alcune tecniche per il Data Mining. Qual è il punto di partenza? Noi abbiamo parlato finora di Data Warehouse, abbiamo parlato di operazioni per l'analisi dei dati, soprattutto orientate a operazioni di tipo OLAP e quindi abbiamo visto le operazioni sull'ipercubo e quindi le operazioni di grid down, roll up, slicing e così via. Quindi operazioni OLAP orientate all'aggregazione dei dati secondo le varie dimensioni di analisi. In realtà quello che vogliamo ancora fare quando abbiamo a disposizione grandi quantità di dati, vogliamo estrarre conoscenza da questi dati. Quindi, in generale, il tema che tratteremo oggi è quello dell'estrazione di conoscenza e quello che andremo a studiare è che cosa vuol dire estrarre conoscenza in modo automatico. Cominciamo a vedere alcuni passi che caratterizzano queste operazioni di estrazione e conoscenza in modo automatico, utilizzando dei dati che sono forniti da varie fonti. I primi passi che sarà necessario fare saranno i seguenti: dovremo selezionare i dati, quindi i dati che saranno di nostro interesse da cui vorremmo estrarre delle informazioni aggiuntive. Ovviamente i dati, come abbiamo visto, per Data Warehouse potranno avere problemi di qualità, quindi sarà necessario fare le operazioni di cleaning per avere una buona base di partenza per estrarre la conoscenza. Successivamente avremo il problema di integrare dati provenienti da fonti diverse. Il passo che dovremmo fare dopo il cleaning è l'integrazione di dati che provengono da fonti diverse e qui dovremmo preoccuparci di fare l'operazione anche di conciliazione, nel caso i dati si riferiscono allo stesso soggetto, quindi tematiche che abbiamo già visto e anche di trasformazione nel caso in cui i formati oppure la struttura delle informazioni siano diverse. Quindi, a questo punto, potremo effettivamente fare le operazioni di trasformazione e strutturazione dei dati risultanti dalle integrazioni di fonti diverse. Tutti questi passi sono i passi che abbiamo sostanzialmente già visto quando abbiamo parlato del caricamento dei dati in un Data Warehouse tramite operazioni di tipo ETM. Quindi, vogliamo successivamente utilizzare questi dati per operazioni di analisi e in questo caso l'analisi sarà di un tipo particolare, quindi vogliamo riconoscere alcune situazioni di tipo notevole che quindi sono nuova conoscenza che viene estratta dai dati che abbiamo raccolto. Tipicamente qui ci andremo a costruire dei modelli del mondo che rappresenteranno delle informazioni relative ai dati che sono stati raccolti. Successivamente dovremo valutare i risultati di questa analisi e, altro passo importante in questa sequenza, è la presentazione dei risultati. In questa lezione, in questo gruppo di lezioni, noi ci concentreremo soprattutto su alcune tecniche di analisi e su come è possibile valutare in modo sistematico i risultati che si sono ottenuti. Una rappresentazione di queste fasi di tipo sintetico è rappresentata in questa figura. Possiamo quindi partire sostanzialmente da quello che abbiamo già raccolto nel nostro Data Warehouse per potere fare le azioni di analisi, abbiamo quindi una operazione di analisi, appunto, abbiamo detto, ci concentreremo sull'estrazione di conoscenza con tecniche di Data Mining, una successiva valutazione e la presentazione. Per poter estrarre la conoscenza, tipicamente partiremo anche da un insieme di informazioni, una base di conoscenze che abbiamo già sul mondo che stiamo esaminando e che possono aiutare a supportare l'ottenimento di informazioni aggiuntive. Per poter fare questa serie di operazioni sono necessarie competenze, in genere, di tipo diverso. Vediamo come, quanto andiamo a studiare si può inserire in un quadro generale, utilizzando quella che è una classificazione delle discipline che è stata proposta da un organismo di standardizzazione. Vediamo qui rappresentata una classificazione di discipline in sotto discipline per quanto riguarda una focalizzazione su quella che viene chiamata la Data Science. Vediamo che qui abbiamo discipline collegate alla gestione di dati e dei processi. Questo è il dominio che noi abbiamo studiato principalmente nell'ambito di questo corso, abbiamo parlato di flussi di dati e dei processi che servono poi a svolgere delle attività, sequenze di attività. Nell'ambito della matematica, della statistica e questo ambito verde di, diciamo, teoria relativa all'analisi dei dati, quindi in ambito matematico, troviamo anche Machine Learning. Andiamo a studiare quelle che sono le tecniche, appunto, matematiche che ci consentono di studiare i dati, le loro correlazioni e di estrarre appunto informazioni a partire dall'insieme di dati. Poi abbiamo un altro ambito che è quello più dell'informatica, quindi lo sviluppo di sistemi software, di sistemi per il supporto all'analisi di algoritmi che vengono resi

efficienti sulla base della teoria matematica che viene definita. Quando consideriamo i vari aspetti, quindi sia quello ingegneristico informatico, sia quello relativo alla gestione dei dati dei processi, sia quello matematico, parliamo in generale di analisi dei dati ma soprattutto ci sono diffuse alcune terminologie. Nella proposta del NIST, viene distinto fra il Data Science e il Data Mining, intendendo come Data Science anche la capacità poi di sviluppare algoritmi più efficienti per riuscire a estrarre informazioni dai dati, usando tecniche matematiche, e distingue il Data Mining come quell'insieme di tecniche che ci consentono di analizzare i dati e i processi con tecniche matematiche. In questo corso, soprattutto andremo a vedere che cosa possiamo fare, quindi ci concentriamo su quali sono gli obiettivi che possiamo raggiungere quando andiamo a utilizzare dei dati per poterne estrarre della conoscenza. Quindi vedremo le categorie di tecniche di Data Mining che sono state proposte, in particolare andremo a vedere quelle basate sul Machine Learning. Un altro aspetto che viene considerato quando si parla di estrazione di informazione dalle grandi quantità di dati è l'aspetto architetturale. Questo è interessante, ovviamente le architetture possono essere di tipo diverso e anche in questo caso abbiamo una proposta da parte del NIST per un'architettura di riferimento. Abbiamo già parlato di architettura di riferimento quando abbiamo parlato delle nostre architetture funzionali, la notazione qui è leggermente diversa, ma vediamo sempre un concetto di flusso. Qui in realtà i flussi sono di tipo diverso, abbiamo delle frecce di colore diverso: le frecce blu rappresentano flussi di dati, le frecce verdi rappresentano l'utilizzo di servizi forniti da altri moduli, le frecce rosse rappresentano il trasferimento di software. Il software che può essere trasferito dai vari partecipanti al processo per poter analizzare i dati. L'ambito in cui è stata proposta questa architettura è un ambito soprattutto relativo a quelli che vengono chiamati i Big Data. Abbiamo visto che possiamo avere diverse fonti di tipo eterogeneo e, in particolare, delle fonti che provengono, ad esempio, da sensori e quindi la disponibilità di grandi quantità di dati che devono essere analizzati e che vogliamo analizzare soprattutto per vedere se da questi possiamo estrarre appunto una conoscenza su quello che è l'andamento dell'organizzazione. Quindi, nell'analizzare i Big Data, noi vediamo due aspetti in questa architettura di riferimento: i flussi che sono necessari per arrivare a quella che è una utilizzazione dei dati da parte di quello che viene chiamato il cliente dei dati. Abbiamo dei fornitori dei dati, in generale abbiamo detto tante fonti che possono essere interne o esterne e dobbiamo svolgere le nostre attività. Come vedete qua abbiamo dei termini che vengono utilizzati per le varie attività, la raccolta dei dati, quella che noi abbiamo chiamato estrazione precedentemente, la preparazione dei dati nei suoi vari aspetti e anche quella che viene chiamata la puration dei dati, il fatto di mantenere questi dati non solo allineati, consistenti e di buona qualità, ma anche annotarli con eventuali altre informazioni, abbiamo prima visto, ad esempio, che possono essere collegate ad informazioni collegate a base di conoscenza, quindi ad esempio etichettare i dati con delle informazioni che si hanno esternamente da questi dati. Invece la parte centrale per noi è quella di analisi e, come abbiamo visto prima, la visualizzazione e anche un'altra tematica è far accedere a questi dati gli utenti. Dal punto di vista delle infrastrutture, ovviamente noi abbiamo l'esigenza di avere diversi strumenti, dal punto di vista dell'infrastruttura a questo punto del corso stiamo ancora parlando del livello applicativo, quindi siamo più interessati allo studio di moduli applicativi che ci consentono appunto di fare operazioni di analisi, ma poi vediamo anche alcuni progetti che abbiamo già incontrato quando abbiamo parlato dei vari livelli di tipo tecnologico. Abbiamo un livello di infrastruttura fisica che può essere fisico oppure di risorse di tipo computazionale o di storage o di tipo virtualizzato. Abbiamo un livello di piattaforma, quindi, in realtà, quello che è necessario in questo caso soprattutto è la capacità di gestire grandi quantità di dati e di immagazzinarli e poi spesso i dati, abbiamo detto, arrivano da diverse fonti e potranno arrivare sia in modalità batch, quella che abbiamo visto finora in cui andiamo a estrarre i dati che sono disponibili o in modo interattivo oppure in modalità streaming, tipica, ad esempio ,di flussi dati provenienti da strumenti di tipo sensori. Quindi, nel poter gestire poi le operazioni di analisi, dobbiamo avere un supporto per poter ricevere i dati e poterli elaborare e memorizzare che diamo per scontato in questa fase. Quindi l'architettura NIST ci fa vedere il processo di elaborazione e ci fa vedere quella che è un'architettura di base per i vari livelli che utilizziamo a livello tecnologico. Abbiamo detto che parleremo soprattutto di Data Mining e apprendimento automatico, quindi dovremo sostanzialmente

ottenere informazioni su condizioni notevoli che si possono identificare valutando i dati. Quindi lo scopo sarà quello di istruire un sistema per estrarre conoscenza. Noi avremo un insieme di dati disponibili da cui appunto vorremmo derivare una nuova conoscenza e soprattutto vorremmo riuscire a capire anche se questa conoscenza effettivamente è una conoscenza di tipo valido, quindi dovremo valutare il risultato. Quindi, in generale, avremo due insiemi di dati da cui partiremo se abbiamo a disposizione dei dati da cui vogliamo poi estrarre della conoscenza. Avremo poi dei dati che chiameremo dati di addestramento, training set, e poi per la valutazione, in genere, utilizzeremo un altro insieme di dati perché vorremmo valutare se questa conoscenza effettivamente è una conoscenza di tipo generale che abbiamo ottenuto, non valida solo per i dati da cui siamo partiti. Qui useremo dei dati in quello che verrà chiamato il test set, quindi dei dati dove potremmo andare effettivamente a verificare se questo risultato, quindi il risultato della conoscenza ottenuto, è un risultato valido testandolo su altri dati. Che tipo di tecniche noi possiamo avere per estrarre conoscenza? Dobbiamo distinguere innanzitutto il modo di addestrare il nostro estrattore di conoscenza. In generale, distingueremo fra due tipologie che verranno chiamate supervisionata, in un sistema di tipo supervisionato, o non supervisionato. Quando abbiamo un sistema supervisionato, definiremo quella che verrà chiamata una variabile target. Quindi vorremmo dire al sistema che nel nostro training set questa variabile assume, per ciascuna delle istanze del training set, un certo valore. Questa operazione verrà chiamata operazione di etichettatura o labelling. Quindi supervisionato vuol dire che ci sarà qualcuno o ci sarà all'interno dei nostri dati una informazione associata alla variabile di target che mi dice qual è il valore che assume la variabile di target in un certo caso. Ad esempio, la variabile di target può essere un'etichetta di classificazione, ad esempio, alto, medio, basso, oppure una classificazione di tipo binario, di tipo sì o no, e quindi noi vorremmo avere nei nostri dati, oltre alle informazioni che sono contenute nei dati, anche dei valori di una variabile target. Nel caso non supervisionato, invece, noi sostanzialmente non avremo questo obiettivo dichiarato, ma sarà l'algoritmo ad identificare delle regolarità, delle caratteristiche tipiche dell'insieme che verrà considerato, quindi andrò ad estrarre regolarità, spesso estrarrò dei gruppi, potrò andare ad identificare anche situazioni particolari oppure delle regole, delle regole generali che si applicano all'insieme che si sta esaminando. Un altro modo di classificare gli algoritmi può essere quello di dividere gli algoritmi in due tipologie: quelle di tipo descrittivo o quelle di tipo predittivo. Descrittivo, come dice il nome, mira a descrivere l'insieme che si sta considerando e quindi, ad esempio, dovrò considerare delle categorie. Questo lo farò tipicamente utilizzando dei dati passati, quindi una storia di dati che ho a disposizione. Nel tipo predittivo, invece, voglio partire dalle conoscenze che ho attualmente per costruire dei modelli che mi consentiranno, quando si presenterà un nuovo elemento della tipologia che sto considerando, di decidere se effettivamente, ad esempio, si verificherà una certa condizione. Un esempio predittivo può essere: sto classificando dei clienti, vorrei sapere se una certa tipologia di cliente andrà o meno a fare un certo acquisto. Noi, per quanto riguarda le tecniche di Data Mining, esamineremo tre tipologie, illustreremo le loro caratteristiche e i metodi per valutarle. Quelle che vengono chiamate regole associative. Io voglio estrarre delle regolarità dal nostro sistema e dire se succede una cosa allora ne succederà un'altra. Tipicamente nel nostro sistema possiamo associare delle regole a delle informazioni. Oppure andremo a vedere le tecniche di classificazione, che mi consentiranno di cercare di capire se, a fronte di certe condizioni, si verificherà poi un certo evento e poi delle tecniche di clustering, in cui voglio raggruppare i dati e capire quali possono essere dei gruppi significativi. Studieremo separatamente, ma quello che possiamo cominciare a definire adesso sono le tipologie che andremo ad avere per queste varie tecniche. Ad esempio, quando vogliamo estrarre delle regole oppure dei gruppi, avremo tipicamente delle tecniche di tipo descrittivo. Quando andiamo a classificare, tipicamente vorremo classificare per poter fare delle previsioni, quindi saranno tecniche tipicamente di tipo predittivo. Un'altra cosa che abbiamo visto che distingue gli algoritmi sono le tipologie di addestramento che possiamo utilizzare e quindi, per le regole associative, noi avremo tipicamente un addestramento non supervisionato, nel caso della classificazione vogliamo addestrare associando delle etichette che mi consentono poi di fare successivamente previsioni su elementi non etichettati, quindi tipicamente avremo tecniche supervisionate o comunque in cui abbiamo una definizione

dei valori per la variabile target e poi, per quanto riguarda il clustering, anche qui vogliamo estrarre della conoscenza, ad esempio, dei gruppi in modo non supervisionato, perché non andremo a definire i nostri gruppi a priori, ma vorremmo estrarli direttamente dai dati.