Co-funded by the European Union

e-EarlyCare-T

Formazione specializzata e aggiornata sul supporto alle tecnologie avanzate per i professionisti e i laureati per l'educazione e per la cura della prima infanzia

MODULO IV.1

Tecniche di osservazione e valutazione adottate da risorse intelligenti: introduzione al Data Mining

Docenti

Dr. Álvar Arnaiz González
Dr. Jose Francisco Díez Pastor
Dra. Sandra Rodríguez Arribas
Dipartimento di Ingegneria Informatica
University of Burgos University of Burgos

e-EarlyCare-T

"Formazione specializzata e aggiornata sul supporto alle tecnologie avanzate per i professionisti e i laureati per l'educazione e per la cura della prima infanzia", e-EarlyCare-T, progetto 2021-1-ES01-KA220-SCH-000032661, è cofinanziato dal programma Erasmus+ dell'Unione Europea, Azione chiave KA220, Cooperazione fra studiosi per Partenariati strategici. Il contenuto della pubblicazione è di esclusiva responsabilità degli autori. Né la Commissione europea né il Servizio spagnolo per l'internazionalizzazione dell'istruzione (SEPIE) sono responsabili dell'uso che può essere fatto delle informazioni qui diffuse".

Indice

I. INTRODUZIONE

Viviamo nella società dell'informazione e della comunicazione, la tecnologia che utilizziamo nel XXI secolo è associata alla raccolta e all'archiviazione di grandi quantità di dati. Il Data Mining (DM) permette di trovare informazioni contenute nei dati che non sempre sono evidenti, poiché, dato il gigantesco volume di dati esistenti, gran parte di questo volume non sarà mai analizzato.

II. OBIETTIVI

1. Conoscere i concetti chiave relativi al Data Mining.
2. Conoscere e applicare semplici tecniche di Data Mining nel campo della cura precoce.

III. CONTENUTI SPECIFICI

3.1. Data Mining

Il Data Mining (DM) è il processo di ricerca e analisi di grandi database per trovare informazioni utili al processo decisionale. Esistono numerose tecniche di DM che impiegano l'analisi matematica per dedurre i modelli e le tendenze esistenti nei dati. In genere, questi schemi non possono essere rilevati con l'esplorazione tradizionale dei dati perché le relazioni sono troppo complesse o perché il volume di dati da analizzare è troppo grande. Attualmente il DM viene utilizzato per l'analisi di grandi quantità di dati in vari campi della conoscenza, come l'istruzione, l'economia, le imprese e l'ambiente.

3.1.1. Concetti di base del Data Mining

Prima di conoscere il processo che viene svolto e i tipi di algoritmi che vengono utilizzati nel DM, è importante chiarire alcuni concetti di base che compaiono frequentemente nella bibliografia associata al DM.

Insieme di dati. È una grande raccolta di dati solitamente organizzata in righe e colonne contenenti variabili e attributi. Ognuno di questi valori è noto con il nome del dato. L'insieme di dati può anche consistere in una raccolta di documenti o file.

Classi o tag. Nel campo del DM, una classe è l'attributo discreto di cui si vuole prevedere il valore in base ai valori di altri attributi. È anche nota come etichetta.

Istanza. Un'istanza è un dato disponibile per l'analisi. Ogni istanza, a sua volta, è composta da caratteristiche che la descrivono. Ad esempio, in un foglio di calcolo, le istanze sono le righe e le caratteristiche delle informazioni memorizzate nelle colonne.

Algoritmo. In informatica, un algoritmo è un insieme di istruzioni definite, ordinate e delimitate per risolvere un problema, eseguire un calcolo o sviluppare un compito. In altre parole, è una procedura necessaria per ottenere un risultato.

3.1.2. Processo di applicazione delle tecniche di Data Mining

Il processo consiste in quattro fasi principali, elencate di seguito:

1. Definizione del problema: è la prima fase in cui un problema specifico viene tradotto in un problema di DM, in cui vengono indicati gli obiettivi dell'analisi e le domande di ricerca.
2. Preparazione e raccolta dei dati. È la fase più estesa del processo, poiché la qualità dei dati è una delle sfide più importanti del DM. I dati grezzi devono essere identificati, puliti e archiviati in un formato predefinito.
3. Modellazione e valutazione: in questa fase vengono selezionate e applicate diverse tecniche di modellazione dei dati (algoritmi), per poi stabilire i parametri e i valori ottimali di queste tecniche.
4. Distribuzione: è l'ultima fase in cui i risultati del DM vengono organizzati e presentati tramite grafici e rapporti. Si veda la Figura 1.

Figura 1. Processo di applicazione delle tecniche di DM.

È importante notare che ogni processo di DM è un processo iterativo, il che significa che il processo non si ferma quando una particolare soluzione viene impiegata. Può essere solo una nuova voce per un altro processo di DM (Rodríguez-Arribas, 2021). In altre parole, in molte occasioni l'applicazione delle tecniche di DM richiede diverse iterazioni e l'uso di diversi algoritmi per poter estrarre i risultati finali della ricerca che stiamo conducendo.

3.2 Tipi di apprendimento nel DM

Esistono numerose classificazioni degli algoritmi utilizzati nel mondo del DM, ma è essenziale capire che esistono due approcci fondamentali: l'apprendimento supervisionato e l'apprendimento non supervisionato. La differenza principale è che nell'apprendimento supervisionato esiste una classe che viene utilizzata per ottenere una funzione che consente di associare i nuovi dati alla classe corrispondente. Nell'apprendimento non supervisionato, invece, non esiste una classe: in questo caso gli algoritmi cercano di scoprire modelli nascosti nei dati senza l'intervento umano sotto forma di tag associati ai dati (Chapelle, Schölkopf y Zien, 2006).

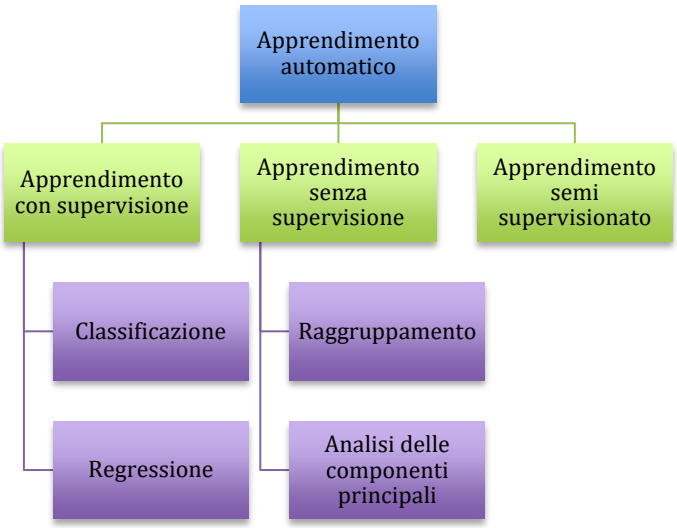

Figura 2. Metodi di estrazione dei dati. Fonte: elaborazione propria.

Quando si deve decidere quale algoritmo utilizzare per l'analisi dei dati, è importante tenere conto del tipo di apprendimento utilizzato, ovvero se si tratta di apprendimento supervisionato o non supervisionato (García, Luengo e Herrera, 2015). A seconda del tipo di apprendimento utilizzato, si utilizzeranno tecniche e algoritmi diversi, come si può vedere nell'immagine precedente.

3.2.1. Apprendimento supervisionato

Una delle modalità di apprendimento del Machine Learning, come già detto, è l'apprendimento supervisionato. L'obiettivo fondamentale dell'apprendimento supervisionato è la creazione di un modello in grado di prevedere i valori corrispondenti agli oggetti in ingresso dopo aver familiarizzato con una serie di esempi, i dati di addestramento. Questa tecnica consiste di due fasi fondamentali:

1. una fase di addestramento in cui si utilizza un insieme di dati etichettati, che contengono i dati di input e i risultati desiderati per quei dati di addestramento con un algoritmo che permette di dedurre una funzione dai dati che stiamo fornendo all'algoritmo.
2. La fase di test, in cui la funzione ottenuta nella fase precedente viene utilizzata per generare nuove previsioni con nuovi set di dati. Vedi la Figura 3.

Il processo è noto come apprendimento supervisionato, poiché conoscendo le risposte di ciascun esempio dell'insieme di addestramento, è possibile correggere la funzione generata dall'algoritmo. L'addestramento dell'algoritmo viene supervisionato correggendo i suoi parametri, in base ai risultati ottenuti iterativamente.

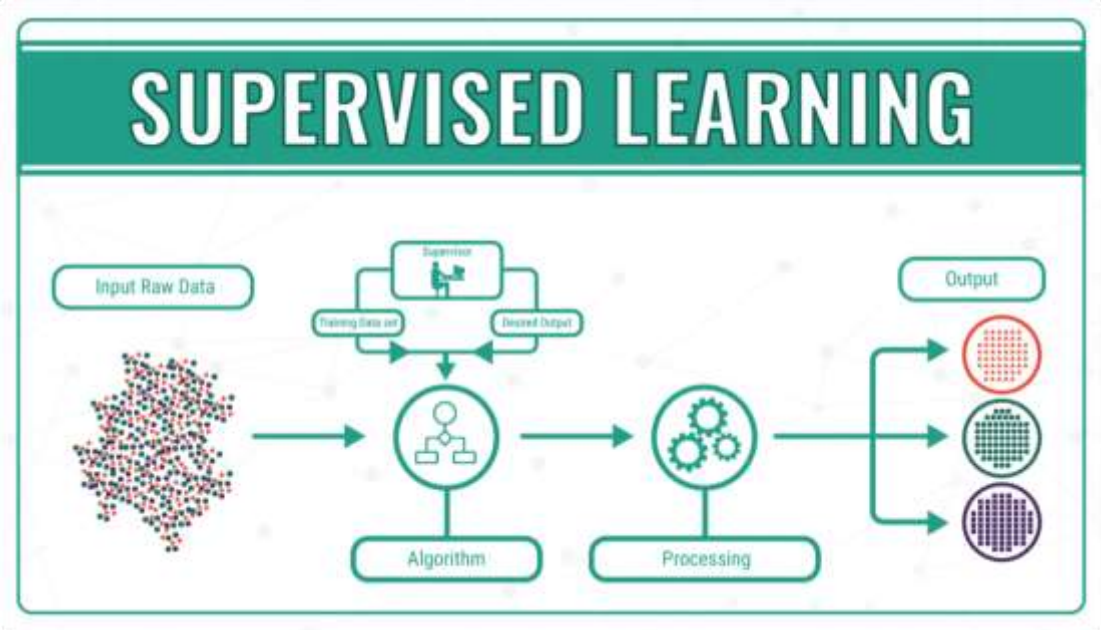

Figura 3. Processo di funzionamento dell'apprendimento supervisionato.
Fonte: ExperiencIA Oracle.

3.2.2. Apprendimento non supervisionato

Questo tipo di apprendimento è l'altro approccio di base al Machine Learning (ML). L'apprendimento non supervisionato prevede dati non etichettati che l'algoritmo deve cercare di capire da solo. L'obiettivo di questo tipo di apprendimento è far sì che la macchina impari senza l'aiuto o le indicazioni dei data scientist, cioè senza supervisione e senza un set di dati di addestramento. Inoltre, la macchina stessa aggiusterà i risultati e i raggruppamenti quando ci saranno risultati più adatti, permettendo alla macchina di capire i dati e di elaborarli nel modo migliore (Figura 4).

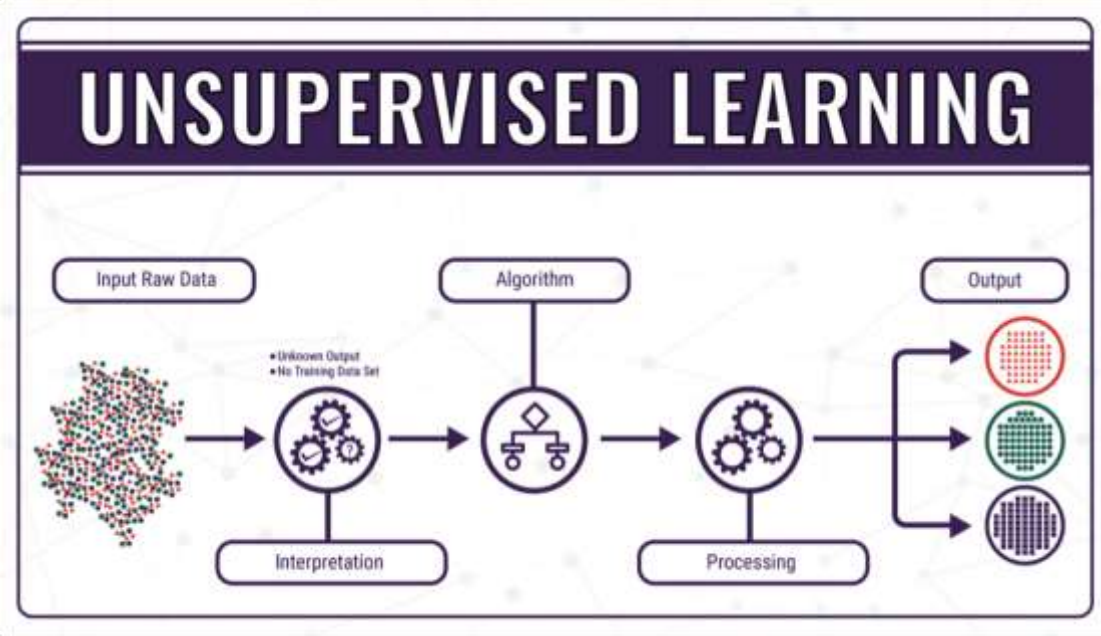

Figura 4. Processo di funzionamento dell'apprendimento non supervisionato.
Fonte: Esperienza di Oracle

L'apprendimento non supervisionato viene utilizzato per esplorare dati sconosciuti e non etichettati. Può rivelare schemi che potrebbero essere stati trascurati o esaminare grandi insiemi di dati che sarebbero troppo impegnativi per una singola persona.

3.2.3 Apprendimento semi-supervisionato

Attualmente sono in corso numerose ricerche sui metodi di apprendimento semi-supervisionato. Queste tecniche di apprendimento automatico utilizzano dati di addestramento sia etichettati che non etichettati: in genere, una piccola quantità di dati etichettati insieme a una grande quantità di dati non etichettati (Zhu e Goldberg, 2009). In altre parole, cercano di migliorare i modelli di previsione ottenuti utilizzando esclusivamente dati etichettati, esplorando le informazioni strutturali contenute nei dati non etichettati. Possiamo dire che l'apprendimento semi-supervisionato cerca di combinare i due approcci tradizionali del DM (apprendimento supervisionato e apprendimento non supervisionato) per mantenere il meglio di ciascuno di essi.

3.3. Algoritmi di classificazione

Gli algoritmi di classificazione sono quelli che utilizziamo quando il risultato atteso è un'etichetta discreta. Sono cioè utili quando la risposta alla domanda di ricerca si trova all'interno di un insieme finito di risultati possibili. Questi algoritmi lavorano generalmente sull'informazione fornita da un insieme di campioni, modelli, esempi o prototipi di addestramento che sono presi come rappresentanti delle classi e mantengono un'etichetta di classe corretta. Questo insieme di prototipi correttamente etichettati è chiamato insieme di addestramento e costituisce la conoscenza disponibile per la classificazione di nuovi campioni. L'obiettivo della classificazione supervisionata è quello di determinare, in base a ciò che si conosce, a quale classe debba appartenere un nuovo campione, considerando le informazioni che si possono estrarre (Figura 5).

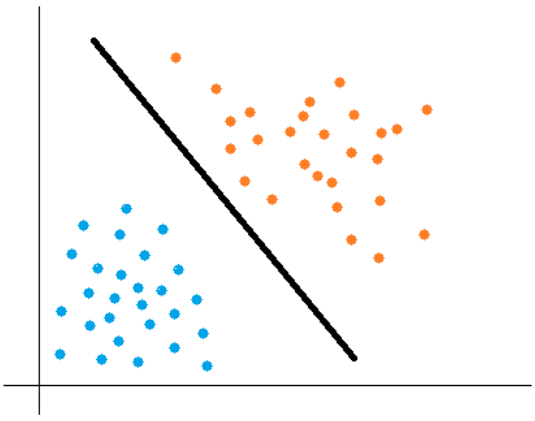

Figura 5. Algoritmo di classificazione. Fonte: elaborazione propria

La classificazione è molto simile al processo di apprendimento delle persone, poiché possediamo la capacità di classificare cibo, libri, animali, pianeti, cioè tutto ciò che ci circonda.

3.4. Algoritmi di clustering

Gli algoritmi di clustering sono responsabili del raggruppamento degli oggetti in un set di dati in base alle loro somiglianze. In questo modo, gli oggetti che si trovano all'interno di un cluster o di un gruppo hanno più somiglianze tra loro che differenze (Figura 6).

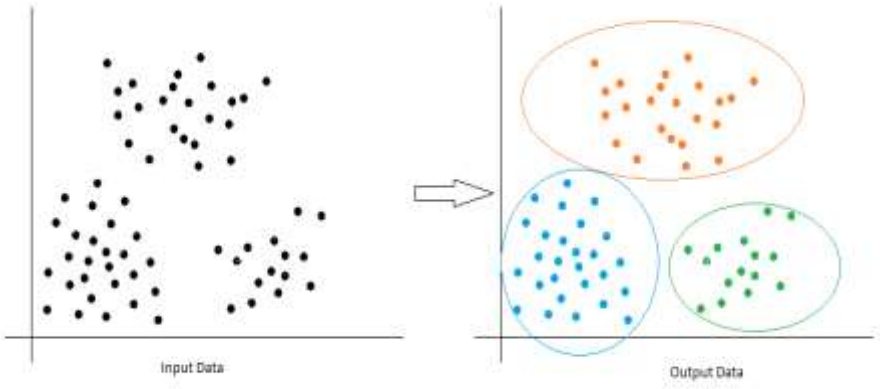

Figura 6. Algoritmo di clustering. Fonte: elaborazione propria

Questi algoritmi lavorano con dati non etichettati, quindi è l'algoritmo stesso che analizza i dati per trovare il numero ottimale di raggruppamenti per l'insieme di dati in ingresso, poiché non abbiamo conoscenze preliminari sulle caratteristiche dei dati e delle loro classi. I raggruppamenti eseguiti dagli algoritmi possono essere di due tipi:

1. raggruppamento rigido: ogni dato appartiene esclusivamente a un gruppo.
2. Raggruppamento morbido (diffuso): i dati possono appartenere a più gruppi in misura diversa, ovvero gli stessi dati possono avere un grado di appartenenza del 60% al gruppo 1 e del 40% al gruppo.

2.3.5. Algoritmi di regressione

Gli algoritmi di regressione sono un sottocampo dell'apprendimento supervisionato il cui obiettivo è stabilire un metodo per la relazione tra un certo numero di caratteristiche e una variabile obiettivo continua. Si tratta di algoritmi che stabiliscono una linea per fornire la tendenza di un insieme di dati; cioè: lo scopo di questi algoritmi è di mettere in relazione un certo numero di caratteristiche e una variabile obiettivo continua (Figura 7).

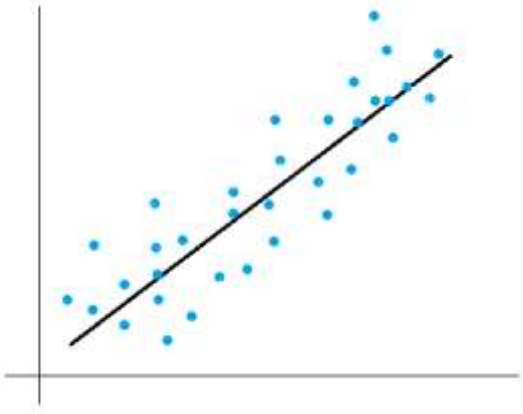

Figura 7. Algoritmo di regressione. Fonte: elaborazione propria

Questa tecnica è utile per prevedere risultati che sono valori continui, il che significa che la risposta alla domanda di ricerca è presentata da una quantità che può essere determinata in modo flessibile sulla base degli input del modello, piuttosto che essere limitata a un insieme finito di etichette come nel caso della classificazione.

3.6. KNIME

KNIME è un'applicazione open-source che può essere usata per lavorare ai nostri set di dati o a set di dati campione:

1. metodi statistici,
2. algoritmi di DM o Machine Learning,
3. tecniche di visualizzazione.

Un software open-source offre molti vantaggi: il suo codice appartiene alla comunità di utenti e sviluppatori, il che garantisce che sarà sempre uno strumento libero e gratuito. Al contrario, il software privato appartiene esclusivamente a un'azienda che può permetterne l'uso gratuito, ma anche far pagare un prezzo elevato o richiedere il pagamento di un abbonamento mensile. È uno strumento progettato per essere semplice da usare. Il concetto più importante nell'uso dello strumento è quello di flusso di lavoro (in spagnolo, workflow). Un flusso di lavoro è una sequenza di passi configurati dall'utente. Formalmente è un insieme di nodi uniti da frecce. Un nodo racchiude diversi lavori che possono essere eseguiti con i dati; ci sono nodi per molti compiti. Un flusso di lavoro potrebbe avere un nodo per caricare un set di dati da un file Excel, quindi un nodo per selezionare gli attributi (colonne) da quel set di dati e un altro nodo per visualizzare le statistiche degli attributi selezionati (Figura 8).

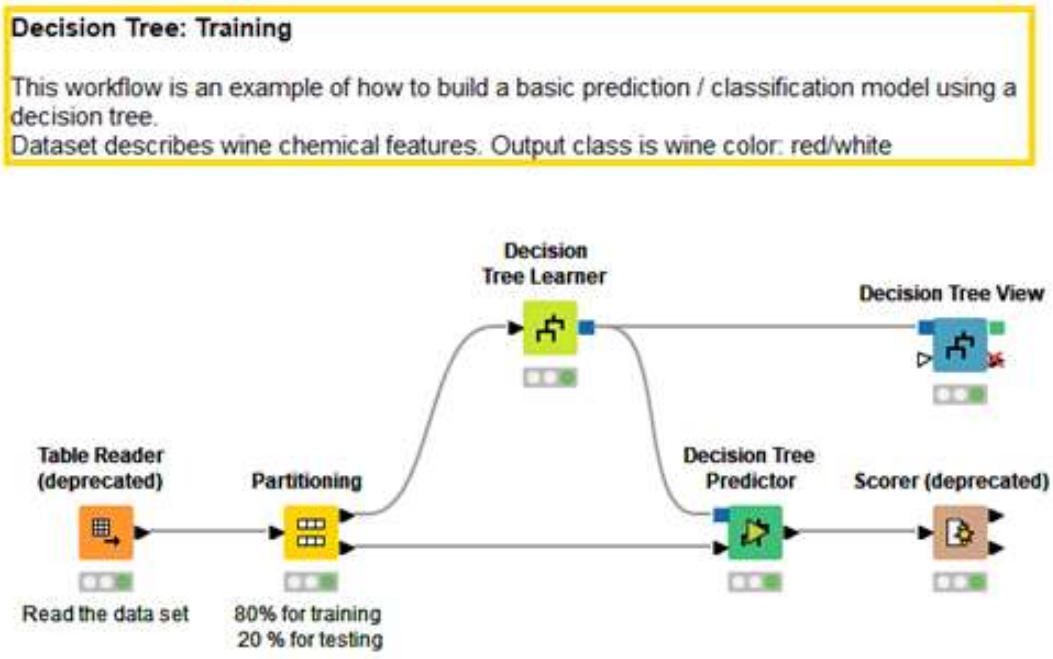

Figura 8. Esempio di flusso di lavoro in KNIME

Le caratteristiche fondamentali per cui KNIME è uno strumento di facile utilizzo sono le seguenti:

1. è uno strumento di "programmazione visiva". L'analisi dei dati può essere effettuata in modo intuitivo, impostando il processo con un semplice clic del mouse. I "nodi" di cui abbiamo bisogno vengono posizionati, senza la necessità di conoscerne il nome o la configurazione, poiché in ogni momento abbiamo un aiuto.
2. Ci sono nodi per applicare qualsiasi procedura o tecnica si voglia, inoltre essendo uno strumento open-source, gli utenti stessi possono creare i propri nodi. Ci sono nodi per:
    a. caricare dati da file o database.
    b. Creare, modificare o eliminare righe o colonne dal set di dati con cui stiamo lavorando.
    c. Calcolare statistiche medie, percentili, correlazioni, ecc.
    d. Combinare dati provenienti da fonti diverse.
    e. Costruire e valutare modelli di apprendimento automatico come classificazione, regressione o clustering.
    f. Visualizzare i dati utilizzando grafici a barre, a torta, a dispersione e altri tipi di grafici più avanzati.
    g. Generazione di report.

3.6.1. Installazione

KNIME è un'applicazione Java, il che significa che è necessario avere installato la macchina virtuale Java prima di poter installare ed eseguire il programma. Per installare il software, dobbiamo andare su https://www.knime.com/downloads; una volta lì, scaricheremo "KNIME Analytics Platform", scegliendo la versione corrispondente al nostro personal computer: Mac, Windows 32 bit (vecchi computer), Windows 64 (computer moderni) o Linux.3.6.2.

3.6.2. L'area di lavoro

L'area di lavoro è la cartella o directory del nostro computer in cui sono memorizzati tutti i progetti realizzati con KNIME. Sarà necessario scegliere un'area di lavoro prima di avviare il programma (si può anche lasciare la cartella che appare di default al momento dell'installazione).

3.6.3. Esempi di utilizzo

Gli esempi sono disponibili nel materiale aggiuntivo, dove vengono rivisti alcuni concetti chiave di KNIME, anche se questi concetti vengono appresi molto meglio se uno studente li esegue sul proprio computer mentre segue le diapositive.

SINTESI

In questa unità sono stati presentati i concetti di base del DM e alcune semplici tecniche di DM da applicare alla ricerca nel campo della cura precoce.

GLOSSARIO

Clustering. È una tecnica di DM, generalmente utilizzata con dati non etichettati, che consente di raggruppare i dati in base alle loro somiglianze o differenze.

DM, Data Mining. È un insieme di tecniche e tecnologie che consentono di esplorare grandi database, con l'obiettivo di trovare modelli ripetitivi che spieghino il comportamento di questi dati e che possano essere utilizzati per trarre conclusioni.

ML, Machine Learning. È una disciplina nel campo dell'Intelligenza Artificiale che conferisce alle macchine la capacità di "apprendere", dall'analisi dei dati, per identificare modelli e supportare il processo decisionale.

BIBLIOGRAFIA

Bogarín, A., Romero, C., & Cerezo, R. (2016). Aplicando minería de datos para descubrir rutas de aprendizaje frecuentes en Moodle. *Revista de Educación Mediática y TIC*, 5(1), 73-92

Chapelle, O., Schölkopf, B. y Zien, A. (2006). *Semi-Supervised Learning: Adaptive computation and machine learning*. MIT Press

Cunningham, P., Cord, M., & Delany, S. J. (2008). *Supervised learning. In Machine learning techniques for multimedia* (pp. 21-49). Springer, Berlin, Heidelberg.

García, S., Luengo, J., y Herrera, F. (2015). *Data Preprocessing in DM* / by Salvador García, Julián Luengo, Francisco Herrera. Springer

Peterson, P. L., Baker, E., & McGaw, B. (2010). *International encyclopedia of education*. Elsevier Ltd

Rodríguez-Arribas, S. (2021). *Minería de datos aplicada al procesamiento automático en el análisis del proceso de enseñanza-aprendizaje* [Tesis doctoral, Universidad de Burgos]. Repositorio académico de la Universidad de Burgos https://riubu.ubu.es/handle/10259/6704

Romero, C., Cerezo, R., Bogarín, A., Sánchez-Santillán, M. (2016). Educational Process Mining: A tutorial and case study using Moodle data sets. En S. Elatia, D. Ipperciel., & O.R. Zaïane (Eds.), DM and Learning Analytics (pp. 3-28). New Jersey: Wiley Online Library. doi: 10.1002/9781118998205

Sáiz-Manzanares, M.C., Marticorena, R., y Arnaiz-Gonzalez, Á. (2022). Improvements for therapeutic intervention from the use of web applications and machine learning techniques in different affectations in children aged 0-6 years. *Int. J. Environ. Res. Public Health*, 19, 6558. https://doi.org/10.3390/ijerph19116558

Sáiz-Manzanares, M.C., Marticorena, R., & Arnaiz, Á. (2020). Evaluation of Functional Abilities in 0–6-Year-Olds: An Analysis with the eEarlyCare Computer Application. (2020). I*nt. J. Environ. Res. Public Health*, 17(9), 3315, 1-17 https://doi.org/10.3390/ijerph17093315

Sáiz-Manzanares, M.C., Marticorena, R., Arnaiz, Á., Díez-Pastor, J.F., y García-Osorio, C.I. (2020). Measuring the functional abilities of children aged 3-6 years old with observational methods and computer tools. *Journal of Visualized Experiments*, e60247, 1-17. https://doi.org/10.3791/60247

Sáiz-Manzanares, M.C., Marticorena, R., Arnaiz-González, Á., Díez-Pastor, J.F., & Rodríguez-Arribas, S. (2019, March). Computer application for the registration and automation of the correction of a functional skills detection scale in Early Care. 13th International Technology, Education and Development Conference Proceedings of INTED2019 Conference 11th-13th (5322-5328). IATED: Valencia. doi: 10.21125/inted.2019.1320

Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1), 1-130.

RISORSE

Software KNIME, https://www.knime.com/downloads