Co-funded by the European Union

e-EarlyCare-T

Specijalizirani trening upotrebom naprednih tehnologija za osposobljavanje i specijalizaciju stručnjaka u području odgoja, obrazovanja i skrbi djece predškolske dobi

MODUL IV.1

Tehnike promatranja i vrednovanja temeljene na korištenju pametnih resursa: Uvod u rudarenje podataka

Nastavnici

Dr. Álvar Arnaiz González
Dr. Jose Francisco Díez Pastor
Dr. Sandra Rodríguez Arribas
Department of Computer Engineering
University of Burgos, Spain

e-EarlyCare-T

Sadržaj

I. Uvod

Živimo u informacijskom, odnosno u komunikacijskom društvu i tehnologija koju koristimo u 21. stoljeću povezana je s prikupljanjem i pohranjivanjem velikih količina podataka. Rudarenje podataka (engl. Data Mining, DM) omogućuje pronalaženje informacija sadržanih u podacima koji nisu uvijek vidljivi, s obzirom na njihovu iznimno veliku količinu. Veliki dio te količine podataka nikada neće biti analiziran.

II. Ciljevi

1. Prepoznati ključne koncepte vezane uz rudarenje podataka
2. Prepoznati i primijeniti jednostavne tehnike rudarenja podataka u području rane skrbi.

III. Sadržaj specifičan za temu

3.1. Rudarenje podataka

Rudarenje podataka (engl. Data Mining, DM) proces je pretraživanja i analize velikih baza podataka kako bi se pronašle korisne informacije u procesu donošenja odluka.

Postoje brojne DM tehnike koje koriste matematičku analizu za identifikaciju obrazaca i trendova koji postoje u podacima. Postojećom analizom podataka, ti se obrasci ne mogu otkriti jer su odnosi među njima presloženi ili je količina podataka jednostavno prevelika za analizu.

Trenutno, rudarenje podataka se koristi za analizu velikih količina podataka u raznim područjima poput obrazovanja, ekonomije, poslovanja, zaštite okoliša, itd.

3.1.1. Osnovni pojmovi u području rudarenja podataka

Prije upoznavanja procesa koji se provodi i vrsta algoritama koji se koriste u DM-u važno je razjasniti neke osnovne pojmove koji se često pojavljuju u literaturi povezanoj s rudarenjem podataka.

Skup podataka

Predstavlja veliki set podataka obično organiziranih u retke i stupce koji sadrže varijable i atribute. Svaka od navedenih vrijednosti ima svoj naziv. Također, skup podataka se može sastojati od više dokumenata ili datoteka.

Klase ili oznake

U području rudarenja podataka, klasa je atribut čiju vrijednost se želi predvidjeti temeljem vrijednosti drugih atributa. Također je poznata kao oznaka (engl. label).

Instanca

Instanca predstavlja svaki podatak dostupan za analizu. Svaka je instanca sastavljena od značajki koje je opisuju. Primjerice, u Excel tablici, instance bi bile redovi (engl. rows), a značajke informacije pohranjene su u stupcima (engl. columns).

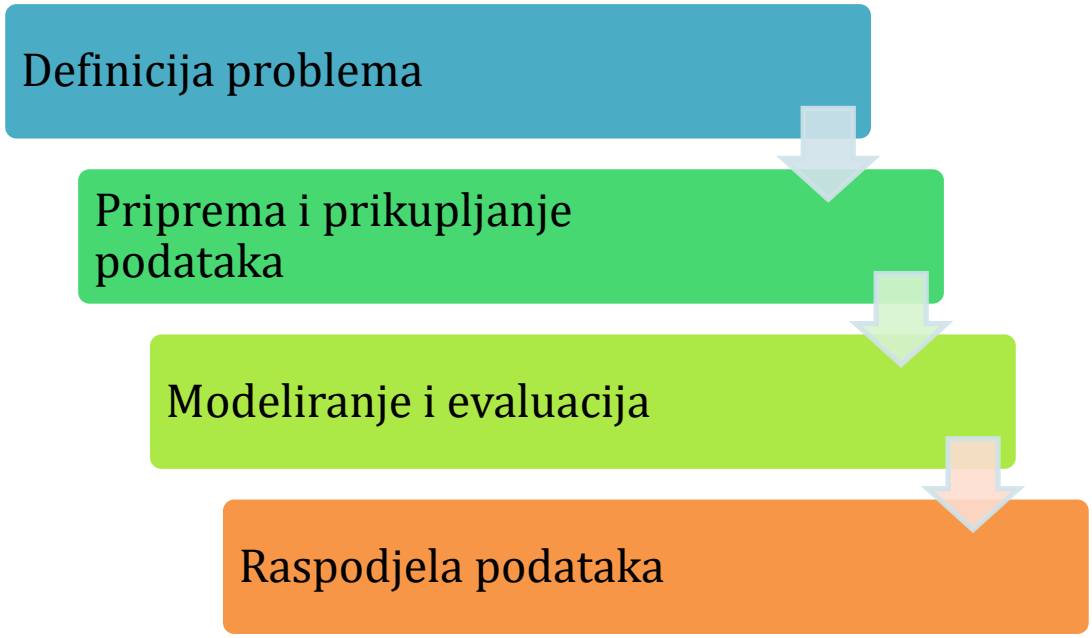
Algoritam

U računalnoj znanosti, algoritam je opisan kao skup definiranih, uređenih i povezanih uputa za rješavanje problema, izvođenje izračuna ili razvoj zadatka. Drugim riječima, radi se o postepenom postupku za postizanje rezultata.

3.1.2. Proces primjene tehnika rudarenja podataka

Proces se sastoji od četiri glavne faze navedene u nastavku:

1. **Definicija problema**: prva faza u kojoj se specifični problem prevodi u problem rudarenja podataka u kojem se zatim postavljaju ciljevi analize i istraživačka pitanja.
2. **Priprema i prikupljanje podataka**: najopsežnija faza procesa jer je kvaliteta podataka jedan od najvažnijih izazova u rudarenju podataka. Neobrađeni („*sirovi*“) podaci se moraju identificirati, „očistiti“ i pohraniti u unaprijed postavljenom formatu.
3. **Modeliranje i evaluacija**: u ovom koraku odabiru se i primjenjuju različite tehnike modeliranja podataka (algoritmi), a zatim se uspostavljaju optimalni parametri i vrijednosti tih tehnika.
4. **Raspodjela podataka**: posljednja faza u kojoj se rezultati rudarenja podataka organiziraju i prikazuju u oblikugrafikona i izvješća.



Slika 1. Proces primjene tehnika rudarenja podataka. Izvor: vlastita obrada

Važno je naglasiti da je svaki proces rudarenja podataka iterativni proces, što znači da se proces ne zaustavlja kada se određeno rješenje implementira. To može predstavljati novi unos za idući proces rudarenja podataka (Rodríguez-Arribas, 2021).

Primjena DM tehnika u mnogim slučajevima zahtijeva nekoliko iteracija i korištenje različitih algoritama kako bi se mogao postići konačni rezultati provedenog istraživanja.

3.2. Vrste učenja u području rudarenja podataka

Postoje brojne klasifikacije algoritama korištene u području rudarenja podataka, no bitno je naglasiti da postoje dva osnovna pristupa: učenje pod nadzorom i učenje bez nadzora. Glavna je razlika u tome što u nadziranom učenju postoji klasa koja se koristi za dobivanje funkcije i koja zatim omogućuje pridruživanje novih podataka odgovarajućoj klasi. S druge strane, u nenadziranom učenju klasa ne postoji. U tom slučaju algoritmi pokušavaju otkriti skrivene obrasce u podacima bez ljudske intervencije u obliku oznaka povezanih s podacima (Chapelle, 2006.).

Slika 2. Metode rudarenja podataka. Izvor: vlastita obrada

Prilikom donošenja odluke koji će se algoritam koristiti za analizu podataka, važno je uzeti u obzir koja se vrsta učenja koristi, učenje pod nadzorom i učenje bez nadzora (García, 2015). Kao što je prikazano na Slici 2., koristit će se različite tehnike i algoritmi ovisno o vrsti učenja.

3.2.1. Učenje pod nadzorom

Kao što je prethodno spomenuto, jedan od modaliteta *strojnog učenja* (engl. Machine Learning; ML) je učenje pod nadzorom.

Temeljni cilj učenja pod nadzorom je stvaranje modela koji ima sposobnost predviđanja vrijednosti putem odgovarajućih ulaznih parametara nakon „stjecanja znanja“ pomoću već unesenih podataka.

Učenje pod nadzorom se sastoji od dva temeljna koraka:

1. Faza učenja u kojoj se koristi skup podataka koji sadrže ulazne parametre i željene rezultate. Osim toga, koristi se i algoritam koji omogućuje izvoz funkcije iz podataka koje mu pružimo.

2. Testna faza, gdje se funkcija dobivena u prethodnom koraku koristi za generiranje novih predviđanja s novim skupovima podataka.

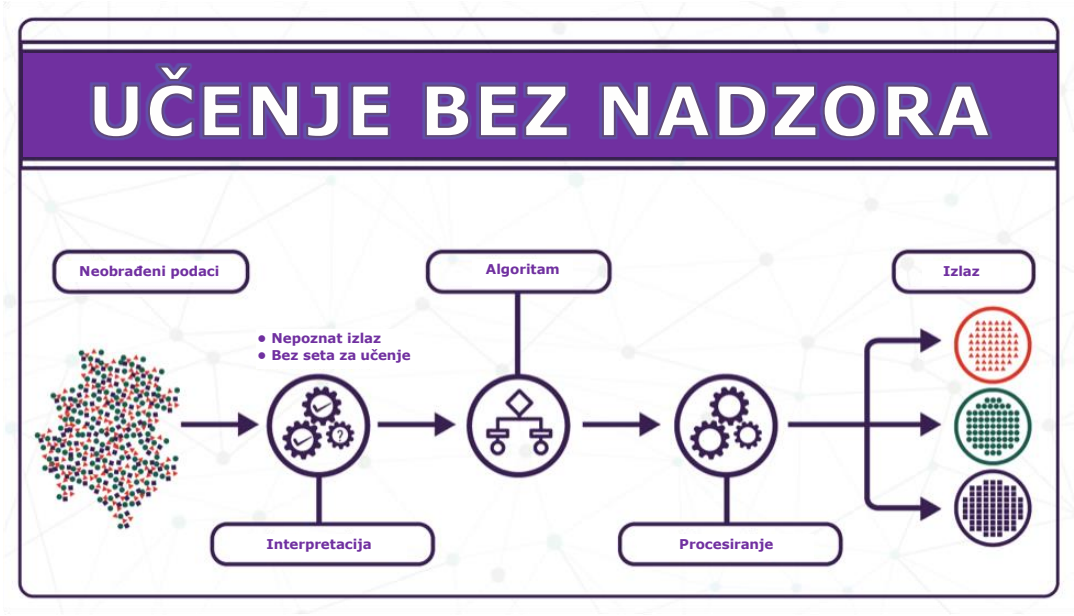
Slika 3:. Prikaz procesa učenja pod nadzorom. Izvor: ExperiencIA Oracle

Proces je poznat kao nadzirano učenje jer je poznavanjem strukture podataka moguće ispraviti funkciju koju generira algoritam. Uvježbavanje algoritma nadzire se korekcijom njegovih parametara, ovisno o iterativno dobivenim rezultatima

. 3.2.2. Učenje bez nadzora

Ova vrsta učenja je drugi osnovni pristup *strojnom učenju* (ML). Učenje bez nadzora sadrži nedefinirane podatke koje algoritam mora sam pokušati razumjeti.

Cilj ove vrste učenja je omogućiti stroju da uči bez pomoći ili uputa dobivenih od strane ljudi, odnosno bez nadzora i bez skupa podataka korištenih za učenje istoga. Pored toga, stroj će sam izvršiti prilagodbu rezultata i njihovo grupiranje kada za to postoje optimalni uvjeti, omogućavajući na taj način stroju razumijevanje podataka i obradu istih najboljim mogućim načinom.



Slika 4. Prikaz procesa učenja bez nadzora. Izvor: ExperiencIA Oracle

Učenje bez nadzora koristi se za istraživanje nepoznatih i nedefiniranih podataka. Može otkriti obrasce koji su promakli iz određenih razlogaili ispitati velike skupove podataka čija je analiza bila prezahtjevna za jednu osobu.

3.2.3. Polunadzirano učenje

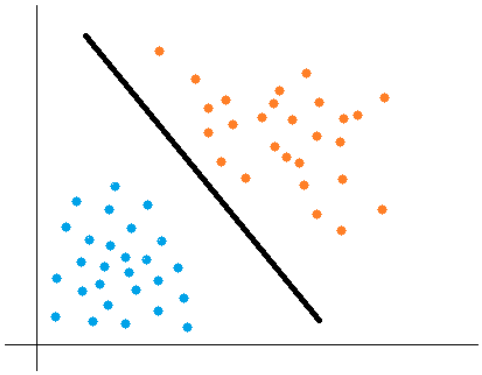
Trenutno se provode brojna istraživanja koja ispituju polunadzirana učenja. Navedena tehnika strojnog učenja koristi manju količinu definiranih i veću količinu nedefiniranih podataka tijekom samog procesa učenja prilikom čega (Zhu i Goldberg, 2009). Polunadziranim učenjem nastoji se poboljšati modele predviđanja koji se dobivaju korištenjem isključivo definiranih podataka, a koji proizlaze iz strukturnih informacija sadržanih u neoznačenim podacima.

Polunadziranim učenjem pokušava se kombinirati dva tradicionalna pristupa rudarenja podataka (učenje pod i bez nadzora) zadržavajući najbolje karakteristike od svakog od njih.

3.3. Klasifikacijski algoritmi

Klasifikacijske algoritme koristimo kada je očekivani rezultat diskretna oznaka. Ova vrsta algoritama korisna je kada se odgovor na istraživačko pitanje nalazi unutar konačnog skupa mogućih ishoda.

Navedeni algoritmi temelje se na informacijama dobivenih iz skupa uzoraka, obrazaca, primjera ili prototipova prethodno provedenog učenja koji predstavljaju određenu klasu prilikom čega zadržavaju njenu oznaku. Skup ispravno definiranih prototipova naziva se set za učenje i to je moguće primjeniti za klasifikaciju novih uzoraka. Po onome što je trenutno poznato, cilj nadzirane klasifikacije je odrediti u koju bi klasu trebao spadati novi uzorak, uzimajući u obzir dobivene informacije.

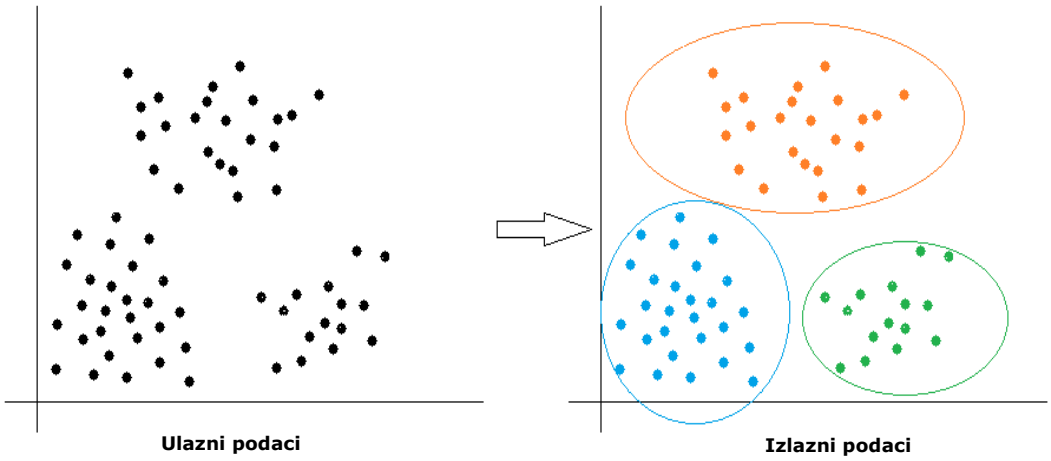


Slika 5. Primjer algorma klasifikacije. Izvor: vlastita obrada

Proces klasifikacije je vrlo sličan procesu učenja ljudi, budući da posjedujemo sposobnost klasifikacije hrane, knjiga, životinja, planeta, odnosno svega što nas okružuje.

3.4. Algoritmi klasteriranja/grupiranja podataka

Algoritmi klasteriranja odgovorni su za grupiranje podataka u svojevrsne skupove, odnosno klastere sukladno njihovim zajedničkim karakteristikama.



Slika 6. Primjer algoritma klasteriranja/grupiranja. Izvor: vlastita obrada

Ovi algoritmi koriste nedefinirane podatke, tako da je sam algoritam taj koji analizira podatke kako bi pronašao optimalan broj i način grupiranja za ulazni skup

podataka budući da nema prethodnog znanja o karakteristikama podataka i njihovim klasama.

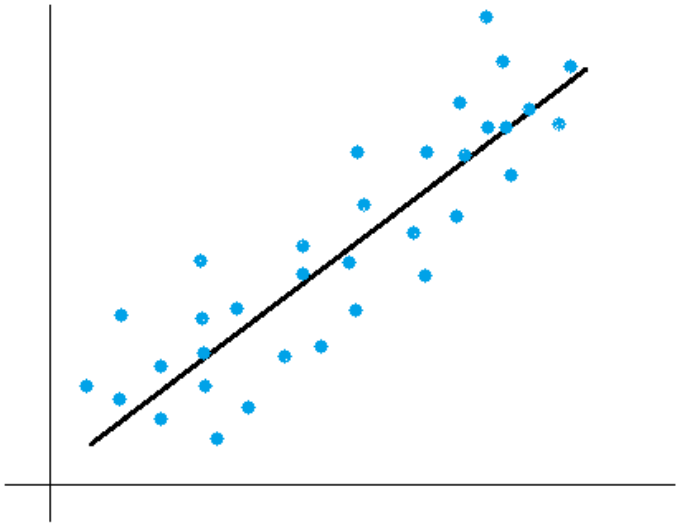
Postoje dvije vrste grupiranja koja izvodi algoritam klasteriranja:

1. **Tvrdi klaster**: svaki podatak pripada isključivo jednoj grupi
2. **Meki (difuzni) klaster:** podaci mogu pripadati više skupina u različitim stupnjevima, odnosno isti podaci mogu imati stupanj pripadnosti, primjerice 60% skupini 1 i 40 % skupini 2.

3.5. Regresijski algoritmi

Regresijski algoritmi su potpodručje učenja pod nadzorom čiji je cilj uspostaviti metodu za odnos između određenog broja karakteristika i kontinuirane varijable.

To su algoritmi koji uspostavljaju liniju za određivanje trenda skupa podataka, odnosno svrha ovih algoritama je povezati niz karakteristika i kontinuiranu objektivnu varijablu.



Slika 7. Prikaz regresijskog algoritma. Izvor: vlastita obrada

Ova metoda je korisna za predviđanje ishoda koji su po svojim karakteristikama kontinuirane vrijednosti, što znači da je odgovor na istraživačko pitanje predstavljen

količinom podataka koja se može fleksibilno odrediti na temelju ulaznih parametara, umjesto da bude ograničena na konačni skup oznaka kao u slučaju klasifikacije, odnosno grupiranja.

3.6. Analitička platforma KNIME

Analitička platforma KNIME je računalni program temeljen na otvorenom kodu koja omogućuje:

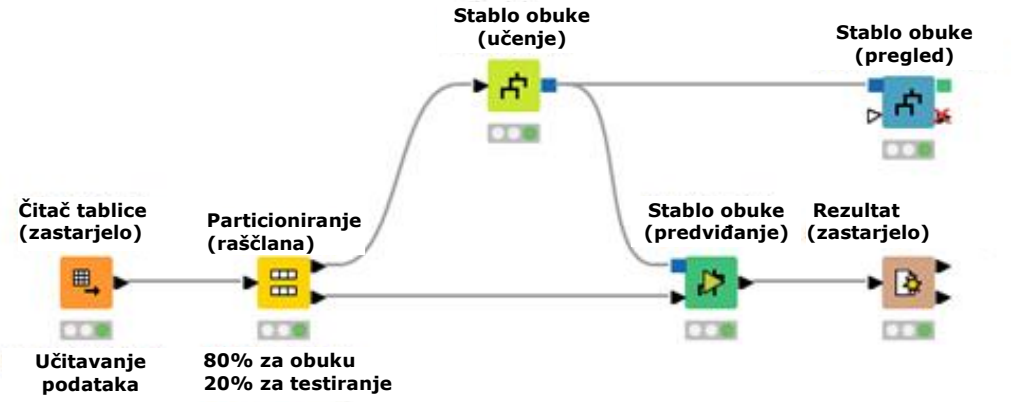
1. Provedbu statističkih metoda
2. Uspostavu algoritama rudarenja podataka ili strojnog učenja
3. Provedbu vizualizacije podataka.

Budući da je računalni program temeljen na otvorenom kodu, ima mnoge prednosti. Njegov osnovni kod „pripada“ zajednici korisnika i razvojnim programerima, što jamči da će uvijek biti besplatan. S druge strane, određeni komercijalni računalni program pripada isključivo tvrtki koja ga je izradila i samo ona može dopustiti njegovo besplatno korištenje, ali i naplatiti visoku cijenu, odnosno zahtijevati plaćanje mjesečne pretplate.

KNIME je platforma dizajnirana za jednostavno korištenje. Ono po čemu je KNIME prepoznat je mogućnost praćenja *tijeka rada* (engl. workflow) tijekom provedbe određene analize ili uspostave algoritama. Tijek rada predstavlja niz koraka koje definira i podešava korisnik. Formalno je to skup spojenih čvorova spojenih. Čvorovi sadrže kodove, odnosno upute za provedbu različitih zadataka koji se mogu izvršiti s postojećim podacima. Tijekom programiranja, korisnik može definirati čvorove za raznorazne zadatke, kao npr.: učitavanje skupa podataka iz Excel datoteke, odabir atributa (varijabli) iz tog skupa podataka, te prikaz statističke odabranih varijabli.

Stablo odlučivanja:

Prikazani tijek rada primjer je generiranja osnovnog modela predviđanja/ klasifikacije koristeći stablo odlučivanja. Baza podataka opisuje kemijske značajke vina. Konačni rezultat je boja vina: crvena/bijela.



Slika 8. Primjer tijeka rada u KNIME-u

Temeljne značajke zbog kojih je analitička platforma KNIME jednostavna za korištenje su sljedeće:

1. Alat za "vizualno programiranje". Analiza podataka može se započeti jednostavnim *klikom* miša. Potrebni čvorovi su već definiranii nema potrebe da korisnik zna detalje njihove uspotave. Ukoliko dođe do određenih problema, korisnik u svakom trenutku ima dostupnu pomoć.
2. Postoje čvorovi za primjenu bilo kojeg postupka ili željene tehnike, osim što je alat otvorenog koda, korisnici sami mogu kreirati vlastite čvorove. Postoje čvorovi za:

    a. Učitavanje podataka iz datoteka ili baza podataka.
    b. Izrada, izmjena ili brisanje redaka ili stupaca iz skupa podataka.
    c. Provedba statističkih metoda: srednje vrijednosti, postotci, korelacije, itd.
    d. Kombiniranje podataka iz različitih izvora.
    e. Izrada i procjena učinkovitosti modela strojnog učenja poput: klasifikacije, regresije ili grupiranja.
    f. Vizualizacija podataka pomoću različitih vrsta grafova i dijagrama.
    g. Izrada izvješća.

3.6.1. Instalacija

Obzirom da je KNIME platforma temeljena na Javi , prije same instalacije i pokretanja iste, potrebno je imati instaliran Java virtualno sučelje.

Platforma je dostupna za preuzimanje putem poveznice: https://www.knime.com/downloads.Na navedenoj web stranici može se preuzeti "KNIME Analytics Platform" odabirom odgovarajuće verzije za osobno računalo: Mac, Windows 32-bit (starija računala), Windows 64-bit (novija računala) ili Linux.

3.6.2. Radni prostor

Radni prostor je mapa ili direktorij korištenog računala u kojem su pohranjeni svi projekti izvedeni s KNIME-om. Prije pokretanja programa, bit će potrebno odabrati gdje će se na računalu nalaziti radni prostor. To može biti i mapa koju je postavio i zadao instalacijski proces platfome.

3.6.3. Primjeri korištenja

Primjeri korištenja platforme su dostupni u dodatnom materijalu gdje se mogu pregledati određeni projekti odrađeni KNIME-om. Valja naglasiti da se koncept upotrebe KNIME-a mnogo bolje savladava ukoliko ih korisnik izvodi na svojem računalu dok prati dodatne materijale.

Sažetak

U cjelini IV.2 obrađeni su osnovni koncepti vezani uz rudarenje podataka i pripadajuća metodologija, koja se može primijeniti u istraživanju u području rane skrbi.

Rječnik pojmova

Grupiranje: tehnika rudarenja podataka koja se općenito koristi s nedefiniranim podacima, što omogućuje grupiranje podataka prema njihovim sličnostima ili razlikama.

DM: rudarenje podataka (engl. Data Mining) je primjena specifičnih algoritama koji omogućuju istraživanje velikih baza podataka, s ciljem pronalaženja ponavljajućih obrazaca koji objašnjavaju uzorak tih podataka i koji se mogu koristiti za donošenje zaključaka.

ML: strojno učenje (engl. Machine Learning), disciplina je u području umjetne inteligencije koja daje računalima mogućnost "učenja".Putem analize podataka pokušava se identificirati obrasce koji definiraju uzorak podataka i omogućitiračunalu donošenje odluka.

Literatura

Osnovna literatura modula

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 Data Mining* / 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.), Data Mining and Learning Analytics (pp. 3-28). New Jersey: Wiley Online Library. doi: 10.1002/9781118998205.ch

Sáiz-Manzanares, M.C., Escolar-Llamazares, M.C., Rodríguez-Media, J. (2019). *Investigación cualitativa. Aplicación de métodos mixtos y de técnicas de minería de datos*. Burgos: Servicio de Publicaciones de la UBU. ISBN: 978-84-16283-79-8.

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-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

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

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

Mrežni izvori

Softver

KNIME

https://www.knime.com/downloads