Universitatea Ştefan cel Mare Suceava

FACULTATEA DE INGINERIE ELECTRICĂ ȘI ȘTIINȚA CALCULATOARELOR

TEZĂ DE DOCTORAT
REZUMAT

CERCETĂRI PRIVIND UTILIZAREA TEHNICILOR ȘI METODELOR DE DATA MINING ÎN CREȘTEREA CALITĂȚII RESURSELOR SISTEMULUI EDUCAȚIONAL LA NIVEL PREUNIVERSITAR ÎN ROMÂNIA

Domeniul Calculatoare și Tehnologia Informației

Coordonator științific,
Conf. univ. dr. ing. **Mirela Danubianu**

Doctorand,
Corina Simionescu

Suceava, 2024

Proiect cofinanțat din Fondul Social European prin Programul Operațional Capital Uman 2014-2020

Această lucrare a beneficiat de suport financiar prin proiectul

Excelență academică și valori antreprenoriale - sistem de burse pentru asigurarea oportunităților de formare și dezvoltare a competențelor antreprenoriale ale doctoranzilor și postdoctoranzilor – ANTREPRENORDOC

***Cod Proiect*:** SMIS 123847
***Cod Contract*:** POCU/380/6/13
***Axa prioritară* 6** - Educație și competențe
Prioritatea de investiții – Îmbunătățirea utilității sistemelor de educație și formare pentru piața muncii, facilitarea trecerii de la educație la muncă și consolidarea sistemelor de educație și formare profesională și a calității lor, inclusiv prin mecanisme pentru anticiparea competențelor, adaptarea programelor de învățământ și crearea și dezvoltarea de sisteme de învățare bazate pe muncă, inclusiv sisteme de învățare duale și de ucenicie.

Apel de proiecte POCU/380/6/13/ „Sprijin pentru doctoranzi și cercetători post-doctorat”

Beneficiar: Parteneri:
Universitatea Ștefan cel Mare Suceava

Cuprins

Cuvinte cheie

Data mining, Metode și tehnici de data mining, Educational Data Mining, Learning Analytics, Machine Learning, Online, Învățarea supervizată, Clasificarea, Măsuri de evaluare a performanței modelelor de clasificare, Algoritmi, Support Vectors Machines (SVM), Arbori de decizie, Random Forest, Naïve bayes, Regresia liniară, Regresia logistică, Învățarea nesupervizată, Rețele neuronale, Deep learning (DL), Text mining, Raționamentul bazat pe cazuri, Analiza sistematică a literaturii de specialitate, RapidMiner, Analiza sentimentelor

1 Introducere

Tehnologia a reorganizat profund modul în care trăim, comunicăm și învățăm. A contribuit la crearea unor stiluri noi de viață, iar rutinele zilnice au devenit mai eficiente. Prin posibilitatea de a ne conecta la nivel global, a redefinit modul în care comunicăm și a revoluționat învățarea oferind oportunități educaționale accesibile și personalizate.

Pandemia COVID-19 a accelerat adoptarea învățării la distanță, școlile și universitățile implementând săli de clasă virtuale. Această experiență a transformat viitorul educației și modul în care instituțiile abordează atât învățarea în format fizic, cât și online.

Data mining (DM) presupune aplicarea unor tehnici capabile să detecteze tipare sau reguli cu ajutorul unor algoritmi potriviți, din cantități mari de date, și reprezintă un pas în procesul complex de descoperire a cunoștințelor din baze de date [1].

În contextul integrării tehnologiilor informaționale în domeniul educației, a fost posibilă generarea, colectarea și stocarea unor volume consistente de date referitoare la aspecte precum obiceiuri de învățare, rezultate obținute etc. Necesitatea analizei acestor date a condus la apariția unei direcții dedicate de descoperire a cunoștințelor din date în care etapa de modelare poartă numele de Educational Data Mining – EDM.

Integrarea modelelor obținute în urma Educational Data Mining (EDM) în procesul educațional are ca scop îmbunătățirea calității educației. Aceasta se poate realiza prin personalizarea strategiilor de predare, organizarea conținutului didactic, construirea unui model de învățare care să se plieze pe nevoile celui care învață și, în general, prin creșterea eficienței de utilizare a resurselor sistemului [2].

1.1 Obiectivele și structura tezei

Obiectivul general constă în realizarea de cercetări privind modul în care modelele rezultate în urma aplicării metodelor și tehnicilor de EDM pot fi utilizate pentru creșterea eficienței resurselor din sistemului educațional la nivel preuniversitar în România. Focusul a fost îndreptat către resursa umană, ca principal factor al calității procesului educațional.

Obiectivul general enunțat este pentru prima dată abordat în România.

Pentru atingerea acestuia, în lucrare sunt propuse următoarele **obiective specifice**:

O1. Studiul aprofundat al conceptului de EDM și al metodelor și tehnicilor caracteristice acestuia.

O2. Realizarea unei analize critice asupra stadiului actual al utilizării EDM în spațiul Uniunii Europene în ansamblu și pentru învățământul preuniversitar.

O3. Constituirea unor seturi de date reale colectate din sistemul de învățământ preuniversitar românesc.

O4. Cercetarea experimentală a modalităților în care utilizarea metodelor și tehnicilor EDM pot servi la îmbunătățirea utilizării resurselor din sistemul preuniversitar din România.

O5. Elaborarea și discutarea unor recomandări privind posibile măsuri de luat pentru îmbunătățirea calității actului educațional în învățământul preuniversitar din România prin eficientizarea utilizării resurselor.

În continuare, teza este structurată astfel:

Capitolul 2 explică conceptul general de descoperire a cunoștințelor din date (KDD – Knowledge Discovery in Databases) pe baza modelelor dezvoltate în diferite contexte de la apariția sa și plasează data mining ca pas central al acestui proces.

În ***Capitolul 3*** sunt descrise metodele și tehnicile utilizate pentru modelarea datelor. Pornind de la ipoteza că modelele de date pot rezolva două categorii largi de probleme – probleme predictive și probleme descriptive – fiecare bazându-se pe tehnici specifice, în acest capitol au fost abordate, în concordanță cu necesitățile ulterioare ale tezei, clasificarea, clusteringul și regulile de asociere.

Capitolul 4 realizează o investigare a stadiului actual privind cercetările și utilizarea EDM în țările membre ale Uniunii Europene, pe baza analizei sistematice a literaturii de specialitate. Am restrâns cercetarea la aceste țări, pentru a plasa educația românească în contextul UE ale cărei norme și directive trebuie să le respecte.

În ***Capitolul 5*** se prezintă seturile de date colectate pentru cercetare. Acestea au fost obținute pe baza completării de chestionare, de către principalii actori din mediul educațional: elevi, profesori și părinți. Sunt șase astfel de seturi de date care au avut ca obiectiv să releve poziția respondenților față de învățământul online în diferite momente. Primele chestionare au fost aplicate în perioada pandemiei Covid-19, în aprilie 2020, când tranziția la școala online a devenit o necesitate. Răspândirea rapidă a virusului a determinat guvernele din întreaga lume să ia măsuri drastice pentru a limita răspândirea, inclusiv închiderea școlilor. Această măsură a afectat milioane de elevi, părinți și cadre didactice, aducând în prim-plan probleme legate de accesul la educație, echitate, sănătatea mentală și dezvoltarea socio-emoțională a copiilor.

Cu scopul de a analiza tendințele, am proiectat și distribuit online chestionare și în 2021, 2022 și în 2024.

Capitolul 6 abordează cercetări experimentale privind utilizarea tehnicilor de clasificare cu scopul de a crea modele care să permită analiza câtorva aspecte importante în creșterea calității proceselor educaționale în sistemul preuniversitar. Adaptarea mediului la modificările de context pornesc de la analiza percepției factorilor implicați, față de situația existentă, de la depistarea acelor influențe care au impact negativ și trebuie minimizate, sau pozitiv și care trebuie potențate.

Capitolul 7 încheie teza marcând concluziile și direcțiile viitoare de cercetare, contribuțiile tezei si un sumar al lucrărilor în care au fost diseminate rezultatele obținute.

1.2 Contribuțiile lucrării de doctorat

Lucrarea sprijină avansul cercetărilor în domeniu printr-o suită de contribuții, atât teoretice cât și practice:

T1: ***Analiză comparativă*** a modelelor asociate procesului de descoperire a cunoștințelor din date, și relevarea caracteristicilor distinctive. (**Capitolul 2**).

T2: ***Analiza tendințelor*** privind utilizarea practică a modelelor KDD, începând cu anul 2007 până în prezent. (**Capitolul2**)

T3: ***Conceptualizarea unor probleme*** din domeniul educațional prin formularea a 23 de întrebări al căror răspuns poate fi furnizat prin explorarea datelor educaționale **(Capitolul 2)**

T4: O investigare a stadiului actual privind cercetările și utilizarea aplicațiilor de EDM în Uniunea Europeană pentru care am realizat o analiză sistematică a literaturii de specialitate (**Capitolul 4**). Procesul de analiză a fost realizat pe baza metodologiei *Kitchenham*, în principalele baze de date academice: *Scopus, ScienceDirect și IEEE Xplore*. Pentru a asigura

relevanța și actualitatea studiului, publicațiile studiate vizează perioada 2013-2023. După aplicarea șirurilor de căutare au fost identificate 847 de studii potențial utile din care, după aplicarea criteriilor de includere și de excludere, au fost selectate 17 studii. Sinteza rezultatelor analizei sistematice a permis obținerea unei imagini clare a stadiului actual al cercetărilor în domeniul EDM la nivelul Uniunii Europene, identificând tendințele majore, lacunele, provocările și oportunitățile pentru viitor.

T5: ***Propunerea unor direcții de cercetare în domeniul EDM în România, ca parte a Uniunii Europene*** (**Capitolul 4).** Concluziile analizei sistematice a literaturii de specialitate privind stadiul actual al Educational Data Mining la nivelul Uniunii Europene, au relevat câteva lipsuri privind cercetările pe alte direcții decât predicția performanțelor funcție de comportamentul de învățare. Pornind de la aceste constatări am propus câteva direcții viitoare de cercetare având ca țintă învățământul preuniversitar din România:

- investigarea influenței factorilor non-cognitivi, precum motivația, atitudinea și implicarea asupra performanței academice;
- investigarea diferențelor de performanță academică între diverse regiuni și școli din România;
- dezvoltarea de modele predictive pentru a anticipa traiectoriile de carieră ale elevilor pe baza performanțelor academice și a intereselor lor;
- utilizarea data mining pentru a identifica elevii cu abilități și talente deosebite și pentru a dezvolta programe de suport pentru aceștia;
- investigarea impactului activităților extracurriculare asupra dezvoltării cognitive și sociale a elevilor.

T6: ***Elaborarea de recomandări*** privind acțiunile necesare de întreprins așa cum rezultă din identificarea factorilor care influențează în mod negativ procesul de învățare (**Capitolul 6)**.

T7: ***Măsurarea consistenței interne a itemilor*** din chestionarele utilizate în teză prin calcularea coeficientului *alpha-Cronbach* folosind software-ul *SPSS*. Valorile obținute sunt cuprinse între 0,736 și 1, ceea ce atestă fiabilitatea și acuratețea instrumentelor utilizate în cercetare (**Capitolul 5)**.

P1: ***Proiectarea a 6 chestionare*** destinate grupului țintă profesori, elevi și părinți și implementarea acestora în *Google Forms.* (**Capitolul 5)**

P2: ***Coordonarea procesului de distribuție a chestionarelor către respondenți*** (**Capitolul 5).** În perioada aprilie 2020 – februarie 2024 au fost distribuite online, la nivel național, 6 chestionare cu scopul de a investiga aspecte privind educația din învățământul preuniversitar din România.

P3: ***Colectarea datelor din chestionare prin preluarea acestora în fișiere Excel și analiza datelor colectate*** din punct de vedere al tipului și calității, precum și al reprezentativității eșantioanelor de respondenți (**Capitolul 5**). Deoarece sondajele au presupus participarea pe bază de voluntariat, se consideră reprezentative eșantioanele formate din răspunsurile a 956 de profesori (2020), 1088 de elevi (2020), 784 de părinți (2020), 7701 de răspunsuri de la cele 3 categorii în anul 2021, 2612 de răspunsuri în 2022 și 1515 răspunsuri culese în 2024.

P4: ***Proiectarea și implementarea proceselor de inducție a modelelor și evaluare a performanțelor, precum și analiza rezultatelor obținute*.** În cadrul cercetărilor efectuate în domeniul *Educational Data Mining*, am implementat procese complexe care implică inducția modelelor, evaluarea performanțelor acestora și analiza rezultatelor obținute. (**Capitolul 6**)

P5: ***Proiectarea și implementarea unor procese de text mining pentru analiza textelor în limba română.*** (**Capitolul 6**)

P6: ***Implementarea principiilor Open Science în cercetarea educațională*****.**
Uniunea Europeană promovează *Open Science* prin directive precum *Planul S* și programele *Orizont 2020* și *Orizont Europa*. Aceste inițiative vizează accesibilizarea și transparența cercetării științifice prin încurajarea în scopul publicării cu acces deschis, partajarea datelor de cercetare conform principiilor FAIR, dezvoltarea infrastructurii digitale și formarea cercetătorilor.
În cadrul tezei mele de doctorat am decis să implementez principiile *Open Science* pentru a crește impactul și vizibilitatea cercetărilor mele. Contribuția mea specifică include următoarele aspecte:

- publicarea datelor cu acces deschis – toate datele colectate în cadrul cercetării mele vor fi postate public pe platforma *Zenodo*, o infrastructură digitală dezvoltată pentru stocarea și partajarea datelor de cercetare; datele vor fi disponibile pentru descărcare și procesare, permițând altor cercetători să valideze rezultatele mele, să extindă cercetarea sau să utilizeze datele în studii conexe;
- conformarea cu principiile *FAIR* (Findable Accessible Interoperable Reusable) – datele vor fi organizate și etichetate astfel încât să fie ușor de găsit, accesat, utilizat în alte platforme, sau contexte de cercetare;
- prin publicarea pe platforma *Zenodo* îmi aduc contribuția la dezvoltarea și consolidarea infrastructurii digitale necesare pentru *Open Science* prin îmbunătățirea accesului la date prin intermediul unor platforme recunoscute la nivel european, aspect ce poate facilita colaborarea internațională și interdisciplinară;
- prin exemplul personal și prin diseminarea rezultatelor și a metodologiei utilizate, urmăresc să inspir și să încurajez alți cercetători să adopte practicile *Open Science*;
- creșterea impactului cercetării prin publicarea datelor cu acces deschis, aspect ce va facilita utilizarea acestora de către un număr mai mare de cercetători, crescând astfel impactul și vizibilitatea cercetărilor mele.

Rezultatele cercetărilor au fost diseminate într-un număr de 12 lucrări. Acestea au fost publicate în jurnale sau în volumele unor conferințe la care au fost prezentate, iar două (A11, A12) au fost prezentate doar în cadrul conferințelor, publicarea făcându-se în volume de rezumate.

Lista articolelor publicate în jurnale și proceeding-uri ale conferințelor:

A1: C. Simionescu, M. Danubianu, D. Marcu, C. O. Turcu, *Online learning after one year of digital schooling in Romania – a survey*, IJCSNS International Journal of Computer Science and Network Security, VOL.21 No.12, December 2021 https://doi.org/10.22937/IJCSNS.2021.21.12.99 - **indexat Web of Science**, **JCI 0.09, Quartile Q4, Accession Number WOS:000755171400005**

A2: Corina Simionescu, Mirela Danubianu, Bogdanel Constantin Gradinaru and Marius Silviu Maciuca, *Educational Data Mining in European Union – Achievements and Challenges: A Systematic Literature Review*, International Journal of Advanced Computer Science and Applications (IJACSA), 15(3), 2024. http://dx.doi.org/10.14569/IJACSA.2024.0150386 **indexat WoS, Quartile Q4, JCI 0.18, Accession Number WOS:001300899000001**

A3: C. Simionescu, M. Danubianu, A.-L. Bârîlă and B. C. Grădinaru, *Factors Influencing Romanian Teachers' Choice of IT Training to Avoid Issues Raised by Online Education: A Data*

Mining Approach, 2024 International Conference on Development and Application Systems (DAS), Suceava, Romania, 2024, pp. 199-204, doi: 10.1109/DAS61944.2024.10541222. **IEEE Digital Library**

A4: **Simionescu, C.,** Marcu, D., & Măciucă, M. S. . (2024). *Toward Better Education Quality through Students' Sentiment Analysis Using AutoML*. BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 15(2), 320-343. https://doi.org/10.18662/brain/15.2/578, (**în curs de indexare WoS**)

A5: Daniela Marcu, Mirela Danubianu, Adina Bărîlă, **Corina Simionescu**, Stefan cel Mare University of Suceava, Romania, *Algorithms for Classifying the Results at the Baccalaureate Exam - Comparative Analysis of Performances*, IJCSNS International Journal of Computer Science and Network Security, VOL.21 No.8, August 2021 http://paper.ijcsns.org/07_book/202108/20210805.pdf,
DOI: https://doi.org/10.22937/IJCSNS.2021.21.8.5 **– indexat Web of Science, JCI 0.09, Quartile Q4, Accession Number WOS:000697025200005**

A6: Maciuca, Marius, Danubianu, Mirela, **Simionescu, Corina.** (2022). *Tendencies in the use of Big Data analytics at a global level*. 155-160. 10.1109/DAS54948.2022.9786116. **IEEE Digital Library**

A7: C. Simionescu, M. Danubianu, D. Marcu (2020), *Analysis of online education romanian schools due to covid-19 pandemics and areas of improvement*, ICERI2020 Proceedings, pp. 3523-3529, 13th annual International Conference of Education, Research and Innovation, Online Conference. 9-10 November, 2020. ISBN: 978-84-09-24232-0 / ISSN: 2340-1095, https://library.iated.org/view/SIMIONESCU2020ANA,
https://doi.org/10.21125/iceri.2020.0787

A8: **Simionescu, C.**, Danubianu, M., & Maciuca, M. S. (2023). How Data Mining and Artificial Intelligence can Contribute to Increasing Academic Performance. Didactica Danubiensis, 3(1), 72–85. https://dj.univ-danubius.ro/index.php/DD/article/view/2467

A9: Corina Simionescu, Mirela Danubianu, Corneliu-Octavian Turcu, *Data mining in educational data – useful tool for sustainable learning development*, EIRP Proceedings of the INTERNATIONAL CONFERENCE European Integration - Realities and Perspectives 16th Edition, 14-15 may 2021, p 349-353, ISSN: 2067–9211 https://dp.univ-danubius.ro/index.php/EIRP/article/view/212/194

A10: Marcu, D., Danubianu M., **Simionescu C.** (2021), *Comparative analysis of predictve models on online education in context of covid-19 – A case study*, INTED2021 Proceedings, pp. 4403-4412, 15th International Technology, Education and Development Conference, Online Conference. 8-9 March, 2021. ISBN: 978-84-09-27666-0 / ISSN: 2340-1079 https://library.iated.org/view/MARCU2021COM, https://doi.org/10.21125/inted.2021.0899

A11: **Simionescu C.,** Danubianu M., Turcu C., *Study on online education in Romania during the Covid-19 pandemic*, 29 june 2021 - 11th International Conference The Danube - Axis of European Identity, Universitatea DANUBIUS Galati

A12: **Simionescu C.**, Danubianu M., Turcu C., Data Mining, *The benefits of data extraction from knowledge in the educational field*, 29 june 2021 - 11th International Conference The Danube - Axis of European Identity, Universitatea DANUBIUS Galati

2 Descoperirea cunoștințelor din date (KDD) și data mining (DM)

Putem spune, cu certitudine, că asistăm la o adevărată ”revoluție informațională”, caracterizată prin procese de generare, colectare și stocare a unor volume imense de date. Paradoxal, aceste volume de date *nu oferă automat* și un volum similar de informații sau cunoștințe [3].

Necesitatea rezolvării acestei situații a condus la apariția și dezvoltarea a noi modalități de procesare a seturilor de date de dimensiuni mari, printre care descoperirea cunoștințelor din date (KDD – *Knowledge Discovery in Data*). Aceasta presupune un proces complex, interactiv și iterativ de identificare a tiparelor noi și potențial utile din date, proces a cărei etapă centrală este data mining. Maturizarea KDD a determinat trecerea de la explorarea retrospectivă la explorarea prospectivă.

Adesea, în mod eronat, se confundă termenul *data mining* cu *descoperirea cunoștințelor în baze de date* [4] [5] [6] [7] [8].

Data mining poate fi definit ca ”explorarea și analiza, prin mijloace automate sau semi-automate, a unei cantități mari de date cu scopul de a identifica tipare utile/relevante” [9];

2.1 Modelarea proceselor KDD

Dezvoltat în mediul academic, dar utilizat pe scara largă în industrie și alte domenii, procesul KDD a făcut obiectul mai multor activități de modelare. Deși proiectate în scopuri diverse, modelele KDD admit faptul că acesta este un proces complex, care se desfășoară iterativ si interactiv. Ele sunt pur conceptuale și independente de sistemele de DM utilizate.

2.1.1 Modelul academic

Modelul academic a fost elaborat de Fayyad. Conform acestui model sunt definite următoarele etape: definirea clară a obiectivelor, obținerea unui set de date țintă, preprocesarea datelor, data mining, interpretarea și verificarea tiparelor obținute anterior (vizualizarea, validarea și interpretarea rezultatelor obținute), după care urmează consolidarea cunoștințelor descoperite pentru susținerea performanțelor sistemului și prevenirea eventualelor anomalii.

2.1.2 CRISP-DM (CRoss-Industry Standard Process for Data Mining)

CRISP-DM este un model utilizat pentru dezvoltarea proiectelor de data mining. Este caracterizat de șase secvențe care pot fi executate ciclic pentru construirea și implementarea unui proces de explorare a datelor.

Nu este obligatoriu ca etapele să se execute iterativ in ordinea prestabilită ci, funcție de rezultatele obținute la un moment dat, se poate face revenire la oricare din pașii anteriori.

Modelul este structurat astfel: înțelegerea procesului/afacerii, înțelegerea datelor, pregătirea datelor, modelarea, evaluarea, implementarea.

2.1.3 SEMMA (Sample Explore Modify Model Assess)

Dezvoltat de Institutul SAS (Statistical Analysis System), propune un proces de explorare a datelor în 5 etape: eșantionare, explorare, modificare, modelare, evaluare.

2.1.4 Studiu comparativ al modelelor KDD

În Tabel 2.1 este prezentat un rezumat al corespondenței între modelele FAYYAD, SEMMA și CRISP-DM.

Tabel 2.1 Rezumatul corespondenței între modelele FAYYAD, SEMMA și CRISP-DM [10]

FAYYAD	**SEMMA**	**CRISP-DM**
Înțelegerea obiectivelor		Înțelegerea problemei
Selecție	Eșantionare	Înțelegerea datelor
	Explorare	
Preprocesarea datelor	Modificare	Pregătirea datelor
Transformarea datelor	Modelare	Modelare
Data mining	Evaluare	Evaluare
Interpretare/Evaluare		Dezvoltare

Se poate deduce că atât etapele procesului SEMMA, cât și etapele CRISP-DM pot fi considerate etape pentru implementarea practică a KDD.

Conform site-ului kdnuggets.com (leader în analiza datelor) - editat de Gregory Piatetsky-Shapiro și Matthew Mayo -, CRISP-DM are cea mai mare pondere de utilizare, în proporție de 43%, conform sondajului realizat în 2014, cu un procent mai puțin față de sondajul realizat în 2007.

Pe baza rezultatelor obținute în urma unui sondaj cu 109 respondenți și furnizate în 2024 de *datascience-pm.com*, CRISP-DM este cel mai utilizat model (49%).

Comparativ cu anul 2014, în 2024 a crescut gradul de utilizare al CRISP-DM cu 6 procente, ajungând de la 43% la 49%. Această creștere indică o adoptare tot mai mare a metodologiei CRISP-DM în rândul profesioniștilor din domeniul data mining și al analizei datelor, subliniind importanța acesteia în procesul de dezvoltare și implementare a proiectelor de data science în ultimul deceniu.

2.2 Data mining – etapă a procesului de descoperire a cunoștințelor din date (Knowledge Discovery în Databases – KDD)

Data mining sau explorarea datelor implică utilizarea algoritmilor capabili să construiască eficient modele (tipare) din cantități mari de date [11].

Contrar ideii că un set de date de dimensiune mare conduce obligatoriu la rezultate bune, s-a demonstrat empiric faptul că nu întotdeauna acest lucru este adevărat. De multe ori este preferabilă eliminarea unor atribute irelevante, redundante sau puternic corelate. Reducerea dimensiunii datelor permite construirea mai rapidă de modele eficiente și, uneori, îmbunătățește acuratețea lor [1].

Data mining constituie elementul central al procesului KDD. Explorarea datelor implică utilizarea iterativă a unor metode specifice, în particular a unor algoritmi specifici, cu scopul de a rezolva două clase majore de probleme: verificarea și descoperirea tiparelor.

Cele mai multe metode de explorare a datelor, cum ar fi clasificarea, gruparea sau regresia, sunt bazate pe tehnicile de învățare automată (machine learning) și pe tehnicile statistice [12].

Tehnicile utilizate pot fi adecvate unui anumit gen de probleme sau de circumstanțe, iar aplicarea lor, în combinație, poate produce rezultate superioare. Optarea pentru o anumită tehnică trebuie să aibă în vedere compatibilitatea dintre problemele care trebuie rezolvate și soluțiile alese.

Cu ajutorul tehnicilor de data mining putem obține cunoștințe care pot fi transformate în acțiune, urmând ca, ulterior, efectul acțiunilor să fie evaluat.

Chiar dacă rezultatele obținute nu sunt pe măsura așteptărilor, trebuie să avem în vedere că atât succesul cât și eșecul pot fi surse de învățăminte pentru acțiuni viitoare, care pot fi orientate, pe viitor, mult mai potrivit.

Toate acestea conturează ideea unui ciclu în utilizarea data mining în cursul căruia se parcurg următorii pași:

- Identificarea obiectivului și a datelor pe baza cărora se poate realiza explorarea;
- Utilizarea tehnicilor potrivite de data mining pentru a extrage informații din colecțiile de date existente;
- Conform informațiilor obținute, se pot lua decizii și întreprinde acțiuni;
- Măsurarea rezultatelor concrete pentru a identifica și alte modalități de exploatare a datelor disponibile.

Aplicarea tehnicilor de data mining poate fi făcută din perspectiva unui demers ascendent sau descendent.

În abordarea descendentă, efortul este orientat spre confirmarea sau infirmarea unor idei (ipoteze) formulate în prealabil prin alte mijloace.

Abordarea ascendentă are o cu totul altă finalitate; ea urmărește extragerea de cunoștințe sau informații noi din datele de care dispunem.

2.2.1 Relația dintre explorarea datelor, învățarea automată și alte domenii de cercetare

Explorarea datelor poate fi plasată la intersecția mai multor domenii actuale de cercetare: statistică, sisteme expert, baze de date, vizualizare, inteligență artificială, învățare automată.

3 Metode și tehnici de data mining

Volumul vast de date generate în diferite domenii, inclusiv în educație, a condus la o creștere a interesului pentru tehnicile și metodele de data mining.

Clasificarea acestora poate fi realizată pe baza mai multor criterii. Unul dintre cele mai importante este considerat a fi tipul de învățare, supervizată sau nesupervizată.

3.1 Învățarea supervizată

Implică utilizarea unui set de date etichetat pentru a antrena modele predictive. Fiecare exemplu din setul de date de antrenament include atât caracteristicile de intrare cât și rezultatele dorite. Parametrii algoritmilor de învățare supervizată pot fi ajustați cu scopul de a minimiza eroarea dintre predicțiile acestora și valorile reale.

În cazul învățării supervizate sistemul este antrenat pentru a rezolva situații noi pe baza situațiilor rezolvate în trecut sau a unui model.

Această categorie include următoarele tehnici:

3.1.1 Clasificarea

Este un proces prin care se construiește (învață, induce) o funcție *f* (clasificator) care asociază fiecărui set de atribute *x* o etichetă de clase *y* dintr-o mulțime predefinită.

Clasificarea datelor este un proces în două etape:

- etapa de antrenare (învățare) în care algoritmul construiește clasificatorul pe baza caracteristicilor;
- etapa de testare în care se estimează precizia regulilor de clasificare.

Pentru stabilirea celui mai bun clasificator se au în vedere câteva criterii: *acuratețea* – capacitatea modelului de a prezice corect eticheta de clasă, *viteza* – timpul necesar configurării modelului, *robustețea* – capacitatea de a prezice corect modelul, chiar și din date *zgomotoase* sau cu valori lipsă, *scalabilitate* – capacitatea unui model de a fi precis și productiv în timp ce manipulează o cantitate tot mai mare de date, *interpretabilitate* - înțelegere facilă, *structura regulilor* – înțelegerea structurii regulilor algoritmilor [15].

În această secțiune sunt prezentate măsuri de evaluare a performanței modelelor de clasificare**:** matricea de confuzie (Confusion Matrix), rata de eroare (Error rate), acuratețea (Accuracy), precizia (Precision), senzitivitatea (Recall) și Scorul F1 (F1 Score).

Există mai multe tipuri de algoritmi de clasificare: regresie logistică (LR-*Logistic Regression),* arbori *(DT Decisin Trees, RF Random Forest),* Lazy *(KNN-K-Nearest Neighbours),* Support Vector Machines *(SVM),* algoritmi Bayes-ieni *(NB-Naïve Bayes) etc.*

În următoarea secțiune se realizează o prezentare generală a algoritmilor frecvent utilizați în clasificare: SVM (Support Vectors Machines), ID3, C4.5, CART pentru DT, Naïve Bayes și Random Forest.

3.1.1.1 Support Vectors Machines (SVM)

Instrument frecvent utilizat pentru analiză predictivă, SVM este o metodă de învățare automată apărută la începutul anilor `90, care își are originile în învățarea statistică.

3.1.1.2 Arbori de decizie

Arborii de decizie constituie o tehnică uzuală pentru construirea modelelor de clasificare sau grupare. Se caracterizează prin reprezentarea tiparelor ca structuri ierarhice simple sub formă de arbore în care fiecare nod frunză indică o clasă de instanțe.

Algoritmul utilizează un criteriu de divizare pentru determinarea celui mai predictiv factor cu scopul de a-l amplasa în rădăcină ca prim punct de decizie în arbore după care se execută căutări de factori predictivi pentru a construi subarborii, până când nu mai sunt date de procesat [18].

În această secțiune sunt descriși cei mai utilizați algoritmi pentru inducția unui arbore de decizie: ***ID3***, ***C4.5*** *și* ***CART***, pseudocodul acestora, avantajele și dezavantajele ID3, Indicele Gini, și este efectuată o analiză comparativă între algoritmii de clasificare ID3, C4.5, CART.

3.1.1.3 Random Forest

Este un algoritm de învățare automată dezvoltat de Breiman în 2001. Este utilizat atât pentru clasificare, cât și pentru regresie. Algoritmul construiește un set care cuprinde un număr mare de arbori de decizie individuali. Fiecare arbore este construit folosind un subset arbitrar de atribute de unde este selectat cel mai bun atribut.

3.1.1.4 Naïve Bayes [19]

Fie setul de date de antrenament D alcătuit din n puncte x_i într-un spațiu d-dimensional și fie y_i clasa pentru fiecare punct, cu $y_i \in \{c_1, c_2, \dots, c_k\}$. Clasificatorul Bayes folosește teorema Bayes pentru a face predicția clasei pentru o nouă instanță de test, x. Se estimează probabilitatea posterioară $P(c_i|x)$ pentru fiecare clasă c_i și se alege clasa care are cea mai mare probabilitate.

În această secțiune se prezintă algoritmul și pseudocodul pentru clasificatorul Naïve Bayes.

3.1.2 Regresia (liniară și logistică)

Regresia este o tehnică asemănătoare cu clasificarea; diferența constă în faptul că în regresie valorile variabilei dependente sunt numerice și continue.

În această secțiune este prezentat sintetic procesul de analiză prin cele două metode și o comparație generală între clasificare si regresie.

Un aspect important al învățării supervizate este calitatea și distribuția datelor, care poate influența semnificativ performanța modelului.

Dezechilibrul de clasă, în care anumite clase sunt subreprezentate în comparație cu altele, poate conduce la modele părtinitoare și performanțe slabe, deoarece algoritmii de învățare automată tind să favorizeze clasele majoritare. Pentru a remedia aceste probleme, echilibrarea datelor devine esențială. Astfel, modelul învață să recunoască și să prezică corect toate clasele.

Tehnici precum *resampling* (*supersampling* și *undersampling*) sunt utilizate pentru corectarea dezechilibrelor. Alegerea tehnicii adecvate depinde de specificul problemei, dimensiunea setului de date și echilibrul dintre clase. Utilizarea combinată a acestor tehnici, în funcție de context, poate duce la obținerea unor modele de învățare automată mai precise și mai robuste.

Supersampling este o tehnică utilă pentru a crește reprezentarea claselor minoritare fără a pierde informații, în timp ce *undersampling* poate simplifica setul de date și reduce timpul de calcul, dar riscă să elimine informații valoroase. În practică, o combinație de tehnici sau

utilizarea unor metode avansate precum *SMOTE* poate adesea oferi cele mai bune rezultate. SMOTE generează noi instanțe sintetice pentru clasa minoritară prin interpolare între instanțele existente [24].

Echilibrarea datelor contribuie la dezvoltarea unor modele mai robuste, mai precise și mai echitabile, care pot oferi rezultate relevante și de încredere în aplicații reale.

3.2 Învățarea nesupervizată

Învățarea nesupervizată se referă la algoritmi care analizează datele fără a utiliza etichete predefinite. Scopul principal este descoperirea de structuri sau modele ascunse în date.

Tehnicile de învățare nesupervizată pot fi utilizate pentru reducerea dimensionalității datelor și identificarea de tipare de comportament în cadrul proceselor de învățare. Reducerea dimensionalității setului de date este un proces de diminuare a numărului de variabile (sau a caracteristicilor) dintr-un set de date, păstrând în același timp cât mai mult din informația relevantă. Acest proces are scopul de a simplifica modelul, a reduce timpul de calcul și a îmbunătăți performanța algoritmilor de învățare automată, în special atunci când se lucrează cu seturi de date de mari dimensiuni sau complexe. Metodele utilizate pot include analiza variației, testul *Chi-pătrat*, informația mutuală, metode de tip *wrapper* (de exemplu, *Recursive Feature Elimination*) și metode de tip *embedded* (de exemplu, *regularizarea Lasso*) [25].

Beneficiile reducerii dimensionalității conduc către îmbunătățirea performanței modelelor prin eliminarea zgomotului și a caracteristicilor irelevante, ceea ce poate îmbunătăți acuratețea și robustețea modelului. În același timp, se reduce complexitatea și timpul de calcul, se previne supraînvățarea și se îmbunătățește interpretabilitatea. Modelele cu mai puține caracteristici sunt mai ușor de interpretat și înțeles, aspect foarte important [27].

3.3 Rețelele neuronale

Rețelele neuronale pot fi utilizate atât în învățarea supervizată, nesupervizată, semi-supervizată cât și în învățarea prin consolidare. Alegerea tipului de învățare depinde de natura problemei și de tipul de date disponibile.

Datorită faptului că rețelele neuronale artificiale se pot plia pe o varietate de probleme, se poate vorbi despre o universalitate a abordării, spre deosebire de metodele clasice de analiză și predicție unde pentru fiecare problemă trebuie identificat un model potrivit [26].

Rețelele neuronale sunt folosite pentru rezolvarea problemelor complexe, modelează date statistice neliniare și operează direct asupra variabilelor numerice. Pentru prelucrarea variabilelor non-numerice este necesară convertirea acestora în variabile numerice.

3.4 Deep learning (DL)

Deep learning (DL) este o metodă de învățare automată care poate fi aplicată în diverse moduri, inclusiv în învățare supervizată, nesupervizată, semi-supervizată și învățare prin consolidare. Alegerea metodei depinde de tipul de date disponibile și de obiectivul specific al aplicației. Este bazată pe arhitecturi de rețele neuronale cu mai multe straturi de unități de procesare, care a fost aplicată cu succes la un set larg de probleme în domeniile recunoașterii imaginilor și procesării limbajului natural, iar în domeniul educațional pentru prezicerea performanței elevilor pe baza activității lor anterioare, detectarea comportamentului nedorit al elevilor, îmbunătățirea transmiterii cursurilor de comunicare vizuală prin transferul de imagini [27].

3.5 Text mining

Text mining este o tehnică modernă de extragere a cunoștințelor din date nestructurate de tip text prin identificarea și explorarea tiparelor interesante în sursele de date. Bazele de date pot conține date în formate diverse: structurate, nestructurate și semi-structurate [29].

Text mining implică extragerea de informații și modele utile din seturi mari de date textuale. Acest proces include de obicei următoarele operații: preprocesare date, reprezentare text, clasificare text, extragere informații, rezumat text, clustering text, vizualizare text [30].

Putem spune că scopul general al text mining este de a transforma textul în date potrivite pentru analiză. Pentru a realiza acest lucru este nevoie de aplicarea algoritmilor de inteligență artificială şi tehnici statistice pentru documente text. Text mining utilizează o gamă largă de sarcini care pot fi combinate împreună într-un singur flux de lucru, în care este posibil să distingem patru diferite etape: recuperarea informațiilor, procesarea limbajului natural (NLP), extragerea informațiilor, explorarea datelor [31].

Algoritmii clasici de clasificare a textului sunt K-Means, SVM și rețele neuronale.

Principala diferență dintre prelucrarea textului și alte tipuri de date constă în faptul că pentru prelucrarea textului este necesară explorarea caracteristicilor înainte de a aplica o anumită tehnică [32].

Clasificarea textului a fost folosită în mediile educaționale pentru diferite scopuri, ca de exemplu: clasificarea automată a activităților în cadrul unui discurs (prelegere a profesorului), discuții și lucru în grup de studenți, clasificarea pe forumuri de discuții etc. [33]

Totuși, cele care se remarcă în acest domeniu sunt: analiza sentimentelor [34], clasificarea întrebărilor [35] și automatic scoring [36].

3.6 Raționamentul bazat pe cazuri

Raționamentul bazat pe cazuri caută răspunsurile (la probleme noi) în experiențele anterioare. Când suntem în fața unei probleme ce poate fi rezolvată cu data mining, putem căuta cazuri similare și să aplicăm rezultatele pentru soluționarea noii probleme. Această metodă poate fi aplicată atât pentru clasificări, cât și pentru predicții. Răspunsul oferit poate fi unul bun pentru mai multe tipuri de probleme.

Cazurile pe care se bazează raționamentul sunt memorate sub formă de înregistrări. Înregistrarea este compusă din setul de atribute care descrie fiecare caz în parte. Putem reprezenta noul caz ca o înregistrare, în care unul dintre câmpuri – cel al cărui valoare trebuie determinată – este vid.

Raționamentul bazat pe cazuri este o tehnică de data mining deosebit de puternică. Prin aplicarea demersului său specific se pot găsi soluții pentru un număr destul de mare de probleme.

Cum data mining prelucrează volume foarte mari de date, calitatea rezultatelor este foarte bună, fiind în strânsă legătură cu acest aspect. O verificare pentru estimarea calității datelor poate fi realizată prin aplicarea tehnicii asupra propriilor date de învățare. Neconcordanțe sau ambiguități pot apărea doar în situația în care numărul înregistrărilor pe care se bazează raționamentul este prea mic.

4 Stadiul actual al cercetărilor și utilizării aplicațiilor de EDM în Uniunea Europeană (UE)

4.1 Introducere

Apărut la începutul anilor 1990, conceptul *sisteme de management al învățării* (Learning Management Systems) - LMS poate fi definit ca o soluție software ce poate facilita desfășurarea procesului educațional acționând ca *hub central* pentru managementul cursului: administrare, comunicare, discuție, crearea de conținut, stocare și evaluarea cursanților [37].

Sistemele de management al învățării (LMS) au fost generate de necesitatea funcționării educației la distanță și au constituit singura modalitate de învățare în România (și nu numai) începând din martie 2020 – momentul debutului pandemiei Covid-19 în țara noastră.

4.2 Educational Data Mining (Explorarea datelor educaționale)

Domeniul Educational Data Mining (EDM) a apărut din necesitatea de a îmbunătăți procesele educaționale prin utilizarea tehnologiilor avansate de analiză a datelor.

În contextul creșterii exponențiale a volumului de date, metodele și tehnicile de data mining sunt importante pentru transformarea datelor brute în cunoștințe utile [38].

Educational Data Mining (EDM) este definit ca ”procesul de dezvoltare și aplicare a metodelor de explorare a datelor pentru a analiza datele educaționale și a descoperi tipare care pot fi utilizate pentru a îmbunătăți procesele educaționale” [39].

Derularea online a activităților de predare-învățare-evaluare generează volume mari de date, care, prelucrate corespunzător, pot conduce la o mai bună înțelegere a elevilor și profesorilor și a condițiilor lor de învățare, la analiza comportamentului de învățare și a factorilor care afectează un studiu de succes, la găsirea unui model asociat performanței învățării și creșterii calității suportului didactic, la îmbunătățirea sistemelor de e-learning, la identificarea motivelor abandonului școlar precum și la luarea deciziilor necesare optimizării managementului tuturor resurselor folosite în sistemele educaționale [40].

Ca urmare, explorarea datelor este o componentă cheie a învățării personalizate și a analizei educaționale.

Procesul EDM se derulează în 4 faze: definirea problemei, colectarea și pregătirea datelor, modelarea și evaluarea modelelor, implementarea [41].

4.3 Analiza sistematică a literaturii de specialitate privind stadiul actual al cercetărilor în EDM la nivelul Uniunii Europene

Cu scopul de a identifica cele mai noi tendințe în domeniul *Educational Data Mining*, am efectuat o analiză sistematică a literaturii (*Systematic Literature Review* – SLR). Pentru a reflecta un stadiu cât mai apropiat de momentul prezent, am ales ca țintă publicațiile din 2013 până în 2023. Este bine cunoscut faptul că țările membre UE au sisteme educaționale cu caracteristici locale specifice și nu există un standard uniform. Cu toate acestea, aplicarea tehnicilor EDM asupra datelor colectate din aceste sisteme poate duce la informații valoroase pentru luarea deciziilor, și de ce nu, la optimizarea acestora la nivelul UE. Acesta este motivul pentru care atenția a fost îndreptată către acele lucrări care au autori afiliați instituțiilor de învățământ din țările membre UE sau care folosesc seturi de date colectate de la aceste instituții.

Noutatea cercetării efectuate constă în faptul că oferă o imagine de ansamblu actualizată asupra cercetărilor și tendințelor în utilizarea EDM în Uniunea Europeană, evaluează nivelul de interes în domeniu și dezvăluie acele aspecte care pot constitui noi direcții de cercetare.

4.3.1 Metodologia de lucru

Am folosit metoda Kitchenham [42] care afirmă că scopul efectuării unui SLR este o analiză amplă a studiilor incluse într-un anumit domeniu pentru a recunoaște lacunele în cercetările existente pentru investigații ulterioare și pentru a oferi o înțelegere cât mai bună a acestui domeniu.

O analiză sistematică a literaturii constă în trei faze distincte necesare unui proces formal de cercetare, fiecare dintre acestea conținând pași și activități specifice.

A. Planificarea analizei sistematice

Am început prin a identifica întrebările de cercetare urmate de o descriere a protocolului predefinit utilizat. Acestea includ informații despre pașii care trebuie parcurși în procesele de analiză sistematică, cum ar fi: strategia de căutare; strategia de selectare a lucrărilor; evaluarea calității; extragerea și sinteza datelor [43].

În plus, un protocol predefinit reduce subiectivitatea cercetătorilor. Pentru a nu duplica informații, elementele de protocol sunt descrise în timpul etapelor în care sunt aplicate.

Scopul și obiectivele analizei. Identificarea întrebărilor de cercetare.

Scopul analizei este acela de a evalua starea actuală a cercetărilor privind metodele și tehnicile EDM și amploarea utilizării acestora în Uniunea Europeană.

Cercetarea bibliografică s-a axat pe lucrări publicate în domeniul EDM, în anii 2013-2023, vizând țări care sunt membre ale Uniunii Europene. Prin identificarea, apoi analizarea și sintetizarea studiilor existente îmi propun să obțin o înțelegere aprofundată a tendințelor de evoluție, a metodelor și tehnicilor utilizate, precum și a rezultatelor obținute în urma aplicării acestora, facilitând astfel înțelegerea progresului, oportunităților și provocărilor actuale.

Obiectivele propuse sunt:

O1: Identificarea metodelor, tehnicilor și algoritmilor utilizați în practica EDM;

O2: Construirea unei imagini asupra EDM prin analiza și sinteza studiilor publicate între anii 2013 și 2023, referitoare la sistemele educaționale din țările Uniunii Europene;

O3: Explorarea beneficiilor adoptării metodelor și tehnicilor de data mining în sectorul educațional;

O4: Identificarea provocărilor, deficiențelor și posibilelor direcții de cercetare viitoare în cadrul EDM.

În conformitate cu obiectivele propuse, am formulat următoarele **întrebări de cercetare**:

RQ1: În ce măsură EDM este implementată la diferitele niveluri ale sistemelor de învățământ din UE?

RQ2: Care este tendința de evoluție a cercetării EDM la nivelul UE?

RQ3: În ce măsură sunt folosite tehnicile de data mining în educație?

Toate acestea conduc la un răspuns documentat la următoarea întrebare sintetică: ***Care este starea actuală a cercetării EDM pentru sistemele de educație din țările UE?***

Criterii de includere și excludere

Pentru a fi sigură că literatura analizată se potrivește cu scopul, obiectivele și întrebările cercetării, am stabilit un set de reguli sub forma criteriilor de includere și excludere. Acestea au permis selectarea lucrărilor relevante, de calitate și adecvate pentru a fi luate în considerare pentru analiza literaturii de specialitate în explorarea datelor educaționale.

Criterii de includere:

I1 : Studii bazate pe autori/seturi de date din sistemul de învățământ al țărilor membre ale Uniunii Europene;

I2: Studii care descriu utilizarea metodelor și tehnicilor de data mining în domeniul educațional al țărilor Uniunii Europene;

I3: Lucrări care abordează în mod consistent subiecte legate de scopul, obiectivele analizei și întrebările de cercetare.

Criterii de excludere:

E1: Cărți și capitole de carte, recenzii de carte, tutoriale, erate, enciclopedii, editoriale;

E2: Studii al căror conținut nu a putut fi accesat prin intermediul contului instituțional oferit de USV;

E3: Studii care reprezintă ele însele o analiză a literaturii de specialitate;

E4: Lucrări scrise în alte limbi decât engleza.

***Metadatele* utilizate în procesul de analiză**

Un aspect important în atingerea obiectivelor propuse este legat de proiectarea setului de date care urmează a fi colectat. Dincolo de citirea articolelor, analiza acestora necesită și extragerea valorilor metadatelor asociate, care pot fi supuse unor procesări diferite. Am luat în considerare următoarele: autorii, titlul lucrării, anul publicării, rezumatul, lungimea lucrării (în pagini), DOI, tipul documentului și cuvintele cheie.

B. Efectuarea analizei

Strategia de căutare și procesul de selecție

Primul pas în acest proces a fost selectarea bazelor de date internaționale adecvate pentru a identifica studiile. Am selectat trei baze de date internaționale: Scopus, Science direct și IEEE *Xplore*. Accesul la bazele de date din această teză a fost asigurat printr-un cont instituțional oferit de Universitatea ”Ștefan cel Mare” din Suceava, România. Astfel, am utilizat portalul *https://www.e-nformation.ro/* care oferă acces la numeroase baze de date în care am consultat resurse științifice.

Procesul de selecție a studiilor parcurge următoarele etape:

- Identificarea inițială a lucrărilor publicate care ar putea satisface în mod plauzibil interogările de căutare;
- Selecția lucrărilor candidat;
- Selectarea studiilor finale de analizat;
- Identificarea studiilor potențial utile.

Faza inițială a implicat parcurgerea unui proces sistematic de identificare și selectare a lucrărilor care corespund scopului cercetării noastre. Pentru a găsi cele mai bune rezultate pentru întrebările de cercetare, am folosit și testat diverse șiruri de căutare.

În Tabel 4.1 sunt prezentate șirurile de căutare din bazele de date considerate, construite după regulile lor specifice.

Tabel 4.1 Șiruri de căutare pentru identificarea studiilor potențial utile

Biblioteca digitală	Șir de căutare
Scopus	(ALL (educational AND data AND mining) AND TITLE (educational AND data AND mining) AND ALL (school)) AND PUBYEAR > 2012 AND PUBYEAR < 2024
Science direct	Educational AND data AND mining AND school Year:2013-2023 Title:educational AND data AND mining
IEEExplore	("All Metadata":educational AND "All Metadata":data AND"All Metadata":mining) AND ("Document Title":educational AND ("Document Title":data AND "Document Title" :mining) AND ("All Metadata":school) Filters Applied: 2013 - 2023

Constatările sunt rezumate în Tabel 4.2.

Tabel 4.2 Studii potențial utile obținute în faza inițială

	Scopus	Science Direct	IEEE
Articole care conțin în toate metadatele șirul ***educational AND data AND mining***	99.731	24.544	8.937
Articole publicate în perioada 2013-2023	91.013	15.356	4.276
Articole care au în titlu șirul de căutare ***educational AND data AND mining***	636	28	183
Articole ce conțin cuvântul ***school*** în toate metadatele	326	16	52
Total	**394**		

În etapa intermediară, ca un prim filtru am citit titlurile, am eliminat duplicatele, am considerat dimensiunea lucrării în număr de pagini ca măsură a coerenței, am analizat lucrările sub 5 pagini și am ajuns la concluzia că o lucrare de mai puțin de patru pagini nu a furnizat suficiente informații. Am cercetat afilierile autorilor și limba în care a fost scrisă lucrarea. Am aplicat filtre suplimentare în șirurile de căutare, acolo unde a fost posibil.

Structura bazei de date Scopus a permis aplicarea de filtre suplimentare, spre deosebire de celelalte baze de date în care am efectuat căutarea. Drept urmare, am aplicat cu ușurință filtre care au condus către reducerea numărului de articole care corespund criteriilor considerate.

Astfel, în **etapa intermediară** am obținut următoarele rezultate: Scopus – 22 de studii, Science Direct – 4 studii și IEEE *Xplore* – 7 studii.

Selecția finală a studiilor

Pentru selecția finală a articolelor de analizat, am efectuat o lectură completă a lucrărilor obținute în etapa anterioară, am evaluat calitatea acestora și le-am păstrat doar pe cele care întrunesc integral criteriile de includere/excludere.

În Tabel 4.3 sunt prezentate rezultatele obținute la finalul fiecăreia dintre cele trei runde de selecție.

Tabel 4.3 Număr de studii selectate pentru analiza finală

Depozitul digital	Studii identificate inițial	Studii selectate (faza intermediară)	Selecția finală
Scopus	326	22	11
Science direct	16	4	2
IEEE *Xplore*	52	7	4
Total	**394**	**33**	**17**

C. Raportare

După cum este prezentat în Tabel 4.3, pentru analiza finală au fost selectate 17 articole.

Pe baza datelor rezumate mai sus, pot fi relevate câteva aspecte interesante despre nivelul actual de implicare a EDM în creșterea calității procesului educațional din Uniunea Europeană.

- Tehnicile de explorare a datelor au fost utilizate în procent de 53% la nivel academic, iar 47% abordează probleme din mediul preuniversitar;
- Educational Data Mining nu numai că facilitează descoperirea eficientă a modelelor și cunoștințelor utile, dar și sprijină luarea deciziilor strategice în acest domeniul educației. Interesul la nivel mondial pentru cercetarea EDM este ridicat și justificat de disponibilitatea tot mai mare a datelor educaționale, dar și de dezvoltarea tehnologiilor;
- În a doua etapă de selecție a studiilor, am remarcat interesul și tendința ascendentă pentru cercetarea EDM în țări precum China, India, Japonia. Numărul de studii selectate în faza finală reflectă un interes mediu al cercetătorilor afiliați instituțiilor de învățământ din țările membre UE. Acest lucru demonstrează necesitatea de a continua cercetările în domeniul EDM. Scopul este de a transforma datele în cunoștințe utile pentru a contribui la îmbunătățirea proceselor educaționale.

Cercetarea include lucrări publicate până în septembrie 2023, deci este posibil ca numărul lucrărilor să crească până în decembrie 2023.

Seturi de date și instrumente

În cele mai multe cazuri, au fost utilizate seturi de date originale. Acestea au fost colectate de la elevi și profesori ca răspunsuri la chestionare sau prin preluarea directă a datelor necesare din LMS-urile utilizate.

Din datele analizate referitoare la instrumentele utilizate rezultă că mediile cele mai utilizate pentru proiectarea și executarea proceselor de data mining au fost *Weka* și *RapidMiner*.

Metodele utilizate cel mai frecvent în explorarea tiparelor din datele din studiile analizate sunt clasificarea urmată de regulile de asociere.

În același timp, studiile analizate au relevat o varietate de provocări legate de calitatea datelor educaționale, confidențialitate și etică, generalizarea rezultatelor și predicție vs. interpretare.

4.4 Concluziile analizei

Originalitatea studiului constă în oferirea unei perspective actualizate asupra stării cercetării și a dinamicii actuale în domeniul Educational Data Mining (EDM) în Uniunea Europeană. Am urmărit evaluarea nivelului de integrare a EDM în sistemele educaționale ale statelor membre ale UE, explorând gradul de utilizare a acestor tehnologii inovatoare la diferite niveluri de educație.

Este evident că nivelul de interes manifestat pentru EDM de către comunitățile academice și educaționale din țările membre UE poate fi îmbunătățit. Pe de altă parte, pentru a contribui la îmbunătățirea calității în educație prin utilizarea tehnicilor EDM, accesul deschis la cercetarea EDM este imperativ.

Lucrarea este structurată în jurul unor întrebări esențiale de cercetare. Prima întrebare investighează măsura în care practicile EDM sunt încorporate în mecanismele educaționale ale Uniunii Europene, relevând nivelul de pătrundere și acceptare a acestor metode în rândul instituțiilor de învățământ.

Faptul că pentru perioada analizată, adică 2013-2023, doar două dintre articole sunt din România, atrage atenția asupra unei prezențe modeste în literatura de specialitate în domeniu la nivel național. Acest lucru evidențiază un potențial neexploatat și oportunitatea de a explora în continuare contribuțiile pe care EDM le poate aduce învățământului preuniversitar românesc, unde adoptarea unor astfel de tehnologii analitice ar putea cataliza progrese semnificative în adaptarea și personalizarea procesului educațional.

Este justificată astfel abordarea de a folosi tehnici EDM cu scopul de a îmbunătăți calitatea învățământului preuniversitar din România.

Cercetările privind utilizarea EDM în Uniunea Europeană din ultimul deceniu s-au concentrat în principal pe învățământul superior. Unul dintre motive ar putea fi acela că la acest nivel LMS-urile sunt utilizate pe scară largă, permițând colectarea de date relativ ușoară.

Interesul pentru aplicarea metodelor EDM în țările membre UE este scăzut. Acest lucru se reflectă, pe de o parte, de numărul mic de lucrări care au îndeplinit criteriile de selecție stabilite, la care au contribuit autori din mai puțin de jumătate din aceste țări, și, pe de altă parte, de tendința cercetărilor publicate în intervalul considerat.

Analiza arată că înțelegerea învățării și a procesului educațional prin tehnicile de explorare a datelor poate oferi o perspectivă profundă și detaliată asupra modului în care elevii învață, se comportă, interacționează și progresează.

Rezumând cercetările efectuate, se poate afirma că tehnicile EDM contribuie la:

- Personalizarea învățării prin identificarea tiparelor de învățare individuale pentru fiecare elev sau pentru grupuri de elevi cu caracteristici similare. Prin utilizarea metodelor și tehnicilor de data mining se pot descoperi preferințele și ritmurile de învățare ale tuturor sau ale grupului. Acest lucru facilitează personalizarea conținutului și a metodelor de predare, asigurându-se că fiecare elev/grup primește sprijinul de care are nevoie.
- Predicția performanței prin dezvoltarea de modele predictive privind performanța academică a studenților. Acest lucru poate ajuta personalul didactic să identifice din timp elevii care ar putea avea dificultăți și să intervină cu măsuri suplimentare.
- Îmbunătățirea proceselor educaționale prin utilizarea tiparelor descoperite în datele despre eficiența metodelor de predare, a materialelor de învățare sau a resurselor educaționale. În acest fel, instituțiile de învățământ își pot îmbunătăți strategiile pentru a asigura succesul experiențelor de învățare ale elevilor.
- Identificarea problemelor structurale prin detectarea și combaterea ratelor ridicate de abandon școlar și a factorilor care contribuie la performanța scăzută a elevilor sau profesorilor. Informațiile ascunse în date pot ajuta la dezvoltarea politicilor educaționale și la alocarea eficientă a resurselor.

Deși există elemente comune între sistemele educaționale la nivel european, particularitățile naționale joacă un rol important în configurarea experienței educaționale.

Astfel, oamenii educați sunt mai conștienți de practicile de sănătate, au rate mai scăzute de morbiditate și mortalitate și, în general, au stiluri de viață mai sănătoase. Concluzionăm că investiția în educație și în cercetare este foarte importantă și absolut necesară.

5 Procesul de generare și colectare a datelor pentru analiza resurselor din învățământul preuniversitar românesc

5.1 Resursele necesare domeniului educațional preuniversitar în România

Școala este factorul principal în realizarea procesului instructiv-educativ, pentru că dispune de ***resursa umană*** specializată. De la începuturile sale (Legea instrucțiunii, 1864), școala românească a fost preocupată de formarea inițială și perfecționarea dascălilor [44].

Familia este una din resursele cele mai importante ale educației, ea fiind partenerul principal al școlii în demersul educațional și în asigurarea succesului în educație.

Un aspect extrem de important și căruia ne propunem să îi acordăm atenție este cel referitor la implicarea emoțiilor în învățare. Emoțiile fixează informația. Putem să povestim efervescent o întâmplare de acum 10 ani, cine a râs, cine a plâns..., dar nu putem să povestim ce am mâncat acum 5 zile la prânz. Dacă am ști, ar însemna să avem 300 de kg! Ne punem întrebarea ce se întâmplă când un copil spune: ”uf, trebuie să învăț!”. Subconștientul execută, dar intervine memorarea de scurtă durată. După ce a terminat de învățat, copilul nu își aduce aminte nimic. Acesta este aportul emoției în învățare și memorare, unul extrem de important!

Începând cu anii 1960, au început să fie dezvoltate diverse teorii despre detectarea şi clasificarea sentimentelor. Analiza sentimentelor ne ajută să înțelegem mai bine care sunt așteptările și nevoile imediate ale elevilor în raport cu educația lor [45].

5.2 Seturile de date utilizate în cercetare

5.2.1 Generare și colectare seturi de date

Începând cu luna aprilie 2020 și până în martie 2024 am proiectat în *Google Forms* și distribuit online 6 chestionare care au avut ca grup țintă profesori, elevi și părinți implicați în sistemul educațional preuniversitar din România, astfel:

C1. Analiza derulării activităților online – profesori (aprilie 2020)

C2. Analiza derulării activităților online – elevi (aprilie 2020)

C3. Analiza derulării activităților online – părinți (aprilie 2020)

C4. Educația online, după 1 an de pandemie – chestionar comun pentru profesori, elevi, părinți (aprilie 2021)

C5. Educația după 2 ani de la începutul pandemiei – chestionar comun pentru profesori, elevi, părinți (aprilie 2022)

C6. Cercetare în scopul contribuirii la îmbunătățirea calității educației din învățământul preuniversitar din România – chestionar comun pentru profesori, elevi, părinți (martie 2024)

Proiectarea acestor chestionare a fost rezultatul colaborării cu specialiști din domeniul științelor educației, psihologiei și managementului.

Pentru distribuirea chestionarelor în mediul online am colaborat cu Uniunea Profesorilor de Informatică din România și cu Inspectoratul Școlar Județean Vrancea, cu mențiunea că prin UPIR distribuirea s-a realizat la nivel național.

Administrarea chestionarelor s-a realizat cu respectarea Legii nr. 677/2001 și a Regulamentului UE 2016/679 al Parlamentului European (GDPR) și nu a implicat din partea respondenților nicio obligație suplimentară. În scopul obținerii de informații sincere, răspunsurile au fost complet anonime, astfel fiind protejată identitatea respondenților.

Am adus la cunoștința respondenților toate aceste aspecte prin secțiunea de descriere a chestionarului.

5.2.1.1 C1. Analiza derulării activităților online – profesori (2020)

Setul de date provine dintr-un chestionar structurat în 24 de întrebări ce poate fi consultat la adresa https://forms.gle/3MKVob1iqKLNbxzL7

Acest chestionar a fost completat de 956 de respondenți.

Fiecare întrebare este concepută pentru a releva diferite aspecte ale experiențelor, competențelor și perspectivelor cadrelor didactice legate de predarea și învățarea online.

Elementele cheie ale setului de date includ: contextul educațional și experiența profesională, nivelul de pregătire și competențe IT, instrumente digitale și resurse educaționale utilizate, percepțiile asupra educației online.

Prin chestionarul distribuit profesorilor ne-am propus investigarea diverselor aspecte legate de educația online, competențele digitale, precum și experiența și percepțiile lor referitoare la predarea și învățarea online. Acest chestionar a fost conceput pentru a colecta date în contextul schimbărilor aduse de pandemia COVID-19 în educație, în special în perioada martie-aprilie 2020.

Descrierea detaliată a întrebărilor, tipurile de răspunsuri (polinomial, binomial, întreg, text), numărul de valori distincte și exemplele de răspunsuri sugerează un set de date complex și bogat, care poate fi explorat pentru a identifica factorii predictivi ai succesului educației online, a percepțiilor profesorilor și a impactului tehnologiei asupra procesului educațional.

O analiză sumară a tipurilor de date colectate din chestionar relevă faptul că acestea îmbracă o mare diversitate de forme, de la date structurate până la date nestructurate.

5.2.1.2 C2. Analiza derulării activităților online – elevi (2020)

Chestionarul pentru elevi a fost conceput cu 21 de întrebări. A fost completat de 1088 de respondenți și se concentrează pe experiențele elevilor față de învățământul online.

Chestionarul poate fi consultat la adresa https://forms.gle/fhgzCUx1SDnxbCfZ6

Setul de date cules permite o analiză profundă a impactului tranziției forțate de la învățământul tradițional la cel online asupra elevilor din diferite medii și cu diverse niveluri de acces la tehnologie. De asemenea, se abordează percepțiile elevilor despre utilizarea tehnologiei în educație, inclusiv gradul lor de confort cu diverse platforme digitale și impactul acestei schimbări asupra procesului lor de învățare.

Datele sunt colectate sub formă de răspunsuri la întrebări specifice, care variază de la cele cu răspunsuri polinomiale și binomiale până la întrebări deschise, oferind astfel o gamă largă de perspective și posibilitatea de a descoperi informații cu grad sporit de subtilitate.

5.2.1.3 C3. Analiza activităților online (părinți – 2020)

În contextul evoluției rapide a tehnologiilor educaționale, înțelegerea dinamicii învățământului online din perspectiva părinților este esențială. Pentru a realiza acest deziderat a fost distribuit online un chestionar adresat părinților cu copii școlari. Acesta explorează diverse dimensiuni ale educației online, accesibilitatea tehnologică și atitudinile părintești față de platformele de învățare digitală. Se concentrează pe diferite aspecte ale educației copiilor, în special în ceea ce privește lecțiile online, accesul la tehnologie și opiniile lor cu privire la eficacitatea și provocările învățării online. Fiecare întrebare este concepută pentru a aduna informații specifice, variind de la date demografice, acces la tehnologie, atitudini față de educația online, până la sugestii de îmbunătățire.

Chestionarul proiectat pentru colectarea opiniilor părinților a fost completat de 784 de respondenți și poate fi consultat la adresa https://forms.gle/Km8WE5QamrYYgXJi7

Acesta a fost conceput cu diverse tipuri de răspunsuri, incluzând răspuns de tip binomial (da/nu), polinomial (multiple opțiuni) și răspunsuri libere, pentru a captura o gamă cuprinzătoare de date. Această abordare combinată permite atât analiza cantitativă a întrebărilor cu răspuns fix, cât și perspective calitative din întrebările deschise. Se pot identifica modele, corelații și diferențe între diferitele grupuri demografice și experiențele lor și atitudinile față de învățământul online.

5.2.1.4 C4. Educația online, după 1 an de pandemie (2021)

Chestionarul *Educația online, după 1 an de pandemie (C4)* se axează pe evaluarea percepției asupra învățământului online din perspective multiple: profesori, elevi și părinți. Acesta a fost distribuit în aprilie 2021, iar abordarea tripartită oferă o perspectivă completă asupra impactului educației online, în perioada imediat următoare pandemiei. Chestionarul poate fi accesat la adresa: https://forms.gle/gtPbrd81FYX4JXzZ6

Întrebările includ [47]: referiri la avantajele percepute ale învățării online, referiri la dezavantajele percepute ale învățării online, scală de tip Likert [48] cu 14 întrebări referitoare la: eficacitatea percepută; atractivitatea percepută; calitatea percepută a evaluării și capacitatea percepută de dezvoltare a abilităților sociale, plus o întrebare directă prin care se evaluează progresul perceput în predarea online în ultimul an.

Chestionarul a fost completat de 7701 de respondenți distribuiți astfel: 942 de cadre didactice, 3224 de elevi și 3535 de părinți.

Setul de date colectat reprezintă nu numai o oportunitate de a examina diferitele perspective legate de tranziția de la învățământul tradițional la cel online, dar și o bază pentru discuții despre eficacitatea învățământului online, adaptabilitatea la tehnologie și diferențele de percepție între profesori, elevi și părinți. Un aspect important este legat de posibilitatea de a identifica factorii care influențează pozitiv sau negativ evoluția educației online.

5.2.1.5 C5. Educația după 2 ani de la începutul pandemiei (2022)

Cu scopul de a evidenția tendințele în evoluția percepției factorilor implicați în procesul educațional am proiectat și distribuit un chestionar și la doi ani de la momentul pandemiei, în anul 2022. Acesta poate fi accesat la adresa https://forms.gle/bgXXv1StDpJ2fdBcA

Chestionarul acoperă o varietate largă de subiecte, de la resursele materiale disponibile în școli, până la percepțiile asupra învățământului online și competențelor digitale. Datele includ atât răspunsuri polinomiale, cât și răspunsuri binare, precum și evaluări pe o scală Likert cu cinci puncte. Setul de date cuprinde răspunsuri la întrebări variate, colectate de la participanți profesori, elevi, părinți, personal didactic auxiliar. Cei 2612 de respondenți au fost distribuiți astfel: 606 de cadre didactice, 708 elevi, 1289 de părinți și 9 didactic auxiliar.

5.2.1.6 C6. Cercetare în scopul contribuirii la îmbunătățirea calității educației din învățământul preuniversitar din România (2024)

Setul de date cules acoperă o gamă largă de subiecte, de la percepții și atitudini până la provocări și nevoi specifice legate de sistemul de educație preuniversitar.

Chestionarul poate fi accesat la adresa https://forms.gle/v6tP5wGXFLyEMi566

Setul de date este compus din răspunsurile a 1515 respondenți la cele 12 întrebări structurate și deschise, adresate unui eșantion divers de participanți (profesori, elevi și părinți). Acesta include variabile polinomiale, binomiale, text, scală Likert cu 5 puncte.

5.3 Evaluarea caracteristicilor seturilor de date

Toate seturile de date au în componență atribute de tip nominal, polinomial, binomial, numeric și text.

Setul de date colectate din **C1**, conține 956 înregistrări caracterizate prin 27 de atribute. Plaja valorilor acestor atribute variază de la două valori (sex, mediu, act_online) până la 183 de valori distincte (instrumente_utiliz).

Chestionarul **C2** a furnizat un set de date cu 1088 de înregistrări, descrise prin 23 de atribute. Aceste atribute au o gamă de valori care pornește de la două valori (acces_net, online vs. clasic) până la 188 valori distincte (platf/aplic_util).

Răspunsurile părinților la chestionarele din grupa **C3** au furnizat un set de date cu 784 de înregistrări, calificate prin 24 de atribute, având, de la două valori distincte (acces_internet, ore_online) până la 189 valori distincte.

Setul de date **C4** a fost gândit în mod unitar pentru cele trei categorii de respondenți, diferențierea fiind făcută prin răspunsul la întrebarea *Vă rugăm să precizați statutul dvs. educațional*. Fiecare exemplu este caracterizat prin 17 atribute, dintre care 14 provin din răspunsurile furnizate pentru componentele întrebării 3, iar două, de tip text, provin din răspunsurile la întrebările 1 și 2. Setul totalizează un număr de 7701 cazuri, structurate astfel: 924 provenite din răspunsurile profesorilor, 3224 din răspunsurile elevilor și 3535 din cele ale părinților. Dacă pentru întrebările cu răspuns de tip polinomial numărul de valori distincte variază între 3 (statut) și 5 (activ_onl_atr_elevi), în cazul atributelor text numărul valorilor distincte este cuprins între 96 (avantaje văzute de profesori) până la 242 (dezavantaje specificate de elevi).

Datele din setul **C5** sunt obținute din răspunsurile oferite după 2 ani de la începutul pandemiei de categorii diverse de persoane implicate în procesul educațional. Setul conține 2612 cazuri, repartizate astfel: 615 cazuri din răspunsurile profesorilor și personalului didactic auxiliar, 708 cazuri din răspunsurile elevilor și 1289 de cazuri provenite din răspunsurile părinților. În afara atributelor de tip text, numărul valorilor distincte este cuprins între 3 și 5. În cazul atributelor text, deoarece răspunsurile au presupus selecție unică dintr-o serie de variante predefinite, acestea prezintă maxim 8 valori distincte.

C6 a fost proiectat în anul 2024 cu scopul de face o analiză a procesului educațional din mediul preuniversitar prin prisma resurselor disponibile și a calității acestora, a satisfacției și a optimismului privind evoluția acestuia precum și a identificării problemelor cu care se confruntă. Ca și în cazul seturilor precedente există date binomiale, polinomiale, numerice și text. În cazul atributelor de tip text numărul valorilor distincte variază între 37 (principale provocări-profesori) și 459 (contribuții la calitatea educației-părinți). Tabelul 5.1 prezintă un sumar al valorilor distincte pentru atributele text din toate cele 6 seturi de date.

Tabel 5.1 Sumarul valorilor distincte pentru atributele text din toate cele 6 seturi de date

Set de date	Categoria de respondenți	Atribute Text	Număr valori distincte	Număr de cazuri
C1	*Profesori*	Avantaje	82	956
		Dezavantaje	86	
		Probleme	926	
		Sugestii	912	
C2	*Elevi*	Avantaje	167	784
		Dezavantaje	87	
		Probleme	489	
		Motiv refuz participare ore online	357	
C3	*Părinți*	Avantaje	134	1088
		Dezavantaje	83	

		Probleme	512	
		Sugestii	457	
C4	*Profesori*	Avantaje	96	942
		Dezavantaje	185	
	Elevi	Avantaje	131	3224
		Dezavantaje	242	
	Părinți	Avantaje	109	3535
		Dezavantaje	227	
C5	*Profesori*	Avantaje online	8	615
		Dezavantaje online	8	
	Elevi	Avantaje online	8	708
		Dezavantaje online	8	
	Părinți	Avantaje online	8	1289
		Dezavantaje online	8	
C6	*Profesori*	Principalele provocări	37	340
		Probleme în educație	251	
		Contribuții la calitatea educației	303	
	Elevi	Principalele provocări	91	600
		Probleme în educație	223	
		Contribuții la calitatea educației	382	
	Părinți	Principalele provocări	69	575
		Probleme în educație	283	
		Contribuții la calitatea educației	459	

Atributele numerice au valori discrete, cuprinse între 1 și 5 sau între 1 și 10, și reprezintă scoruri acordate pentru diferite aspecte cum ar fi: autoaprecierea competențelor didactice, utilitatea sau eficiența aplicațiilor online, etc. Este interesant de studiat media acestor valori (semnificând scorul mediu primit de fiecare caracteristică în parte) pentru a avea o imagine generală a percepției legate de aceste caracteristici.

Faptul că volumul datelor colectate nu este foarte generos pentru procesări prin data mining, a determinat o investigație privind posibilitatea folosirii tehnicii de învățare prin transfer (transfer learning). Aceasta presupune utilizarea de modele pre-antrenate pe date dintr-un domeniu sursă similar celui analizat și aplicarea acestora pe datele colectate, cu scopul de a contribui la atenuarea dificultăților legate de cantitatea și calitatea limitată a seturilor de date disponibile și reducerea timpului și costurilor necesare producerii unui număr suficient de date etichetate manual.

5.4 Considerații etice

În această secțiune sunt descrise considerațiile etice pe care le-am avut în vedere în contextul realizării acestei cercetări: confidențialitatea și protecția datelor, consimțământul informat, echitatea și non-discriminarea, transparența și responsabilitatea în interpretarea rezultatelor, respectarea drepturilor de proprietate intelectuală și reutilizarea datelor.

5.5 Contribuții

Pe parcursul acestui capitol a fost abordat procesul de generare și colectare a seturilor de date utilizate ulterior în analiză. Din gama largă de resurse pe care se bazează sistemul educațional am ales capitalul uman ca factor esențial în creșterea calității acestui sistem.

Se pot remarca următoarele contribuții:

- proiectarea și implementarea în *Google Forms* a 6 chestionare care au fost distribuite profesorilor, elevilor și părinților din mediul preuniversitar din România;
- coordonarea procesului de distribuire a chestionarelor către respondenți;
- colectarea datelor din chestionare;
- analiza datelor colectate cu scopul evaluării calității seturilor obținute.

6 Cercetări experimentale privind utilizarea metodelor și tehnicilor de data mining pentru creșterea calității resursei umane din sistemul de educație

O perspectivă amplă asupra strategiilor și obiectivelor educaționale din România pentru perioada 2021-2027 ne este oferită în documentul *Reforme în derulare și evoluții în materie de politici* [50] publicat pe site-ul comisiei europene. Sunt detaliate reformele din diferite sectoare ale educației, inclusiv educația timpurie, profesională și învățământul superior, precum și dezvoltarea abilităților transversale. Se subliniază importanța unui cadru strategic pentru îmbunătățirea sistemului educațional, conform obiectivelor europene, și consideră proiectul ”România Educată” ca fiind un reper major pentru reforma educației. De asemenea, se conturează direcții strategice pentru inovație și modernizare. Acestea includ digitalizarea procesului educațional, adaptarea curriculumului pentru a dezvolta competențe relevante în era digitală și formarea profesională continuă a cadrelor didactice.

6.1 Analiza sentimentelor / opiniilor profesorilor, elevilor și părinților cu privire la desfășurarea orelor online în perioada pandemică

Obiectivul a fost acela de a **construi modele de analiză a sentimentelor tuturor categoriilor de persoane implicate în educația preuniversitară din sistemul românesc, față de diferitele aspecte ale procesului de educație online**. Pe de o parte aceste modele servesc pentru viitoare predicții, pe de altă parte sunt și o modalitate prin care se pot evidenția și conștientiza acei factori care au conotație negativă sau pozitivă în contextul acestor modele, ca bază pentru acțiuni coerente de adaptare.

Construirea unor modele prin care să fie analizate sentimentele profesorilor, elevilor și părinților cu privire la ***desfășurarea lecțiilor online în perioada pandemică***, a constituit primul pas al cercetărilor aplicative.

Am căutat răspuns următoarelor întrebări:

- **I1**: *Dată fiind structura complexă a setului de date care cuprinde atât caracteristici structurate cât și nestructurate, care dintre aceste tipuri oferă modele cu performanțe superioare?*
- **I2**: *Ce tehnici de data mining sunt mai eficiente?*

Au fost utilizate cele trei seturi de date generate din răspunsurile date de profesori, elevi și părinți la chestionarele aplicate în timpul pandemiei. Spre a răspunde întrebării ***I1***, pentru fiecare set de date considerat, au fost folosite la început datele colectate din răspunsurile structurate, apoi doar caracteristicile colectate din răspunsurile libere, deoarece acestea surprind elemente mai subtile ce pot descrie o anumită clasă, iar în final setul complet de date.

Răspunsul la întrebarea ***I2***, a presupus analiza celor mai bune modele rezultate din ***I1*** și raportarea celor mai eficiente tehnici.

Deoarece nu există un item prin care respondenții să specifice ce sentimente au față de forma de învățământ online, un pas important și, probabil cel mai dificil în pregătirea procesului de creare a modelului a fost cel de ***etichetare a setului de date***. A fost considerat întregul set de date și a participat un colectiv format din trei specialiști. În urma unei analize preliminare au fost depistate caracteristici ale căror valori sunt aproape constante sau care au foarte multe valori distincte deci nu pot contribui în măsură semnificativă la stabilirea etichetei. În contrast,

caracteristicile cu răspuns liber (avantaje, dezavantaje, probleme, sugestii) furnizează elemente care pot fi indicii privind sentimentele celor chestionați.

Etichetarea finală a fost rezultatul unui proces iterativ, executat în mai mulți pași.

O analiză a claselor (pozitiv/negativ) pentru cele trei seturi utilizate a condus la concluzia că pentru seturile de date care au fost constituite din răspunsurile profesorilor și elevilor clasele sunt relativ echilibrate, în timp ce pentru setul de date colectat din răspunsurile părinților clasele sunt dezechilibrate.

Dezechilibrul claselor *pozitiv/negativ* în răspunsurile părinților indică faptul că răspunsurile părinților nu sunt distribuite uniform între categorii, existând o predominanță semnificativă a clasei *pozitiv* în detrimentul clasei *negativ*. Acest dezechilibru poate avea diverse implicații în analiza datelor deoarece un set de date dezechilibrat poate influența performanța algoritmilor de data mining, deoarece aceștia pot fi părtinitori spre clasele majoritare, ignorând sau subevaluând clasele minoritare.

În cele ce urmează, am efectuat cercetări privind posibilitățile de echilibrare a claselor și de creștere a performanțelor modelării utilizând metode de echilibrare prin supraeșantionare și subeșantionare, dar și metode *wrapper*.

6.1.1 Pregătirea și modelarea datelor structurate

Datele structurate sunt datele colectate din componenta cu răspunsuri scurte, predefinite a chestionarelor. Procesul de preprocesare se realizează prin acțiuni efectuate atât asupra datelor din fișierul Excel cât și prin operatori specifici RapidMiner asupra datelor importate în Repository și are în componență următoarea succesiune de pași:

- selecția atributelor provenite din întrebări cu răspuns predefinit;
- înlocuirea denumirilor caracteristicilor (care în setul original sunt reprezentate prin propoziții lungi) cu sintagme sugestive, dar scurte;
- transformarea valorilor atributelor astfel încât, păstrând semnificația să se reducă numărul de valori (ex. elev IX și elev XII au fost înlocuite cu „liceu”; similar am procedat pentru elevii de gimnaziu care s-au transformat în „gimnaziu”);
- discretizarea atributelor numerice și transformarea lor în atribute descriptive (scorurile cu valori întregi au fost transformate în calificative de tip foarte bine, bine, satisfăcător sau nesatisfăcător);
- completarea valorilor lipsă și stabilirea atributului etichetă.

Pentru realizarea transformărilor în mediul RapidMiner Studio 10.3 am proiectat, implementat și executat procesul din Fig. 6.1.

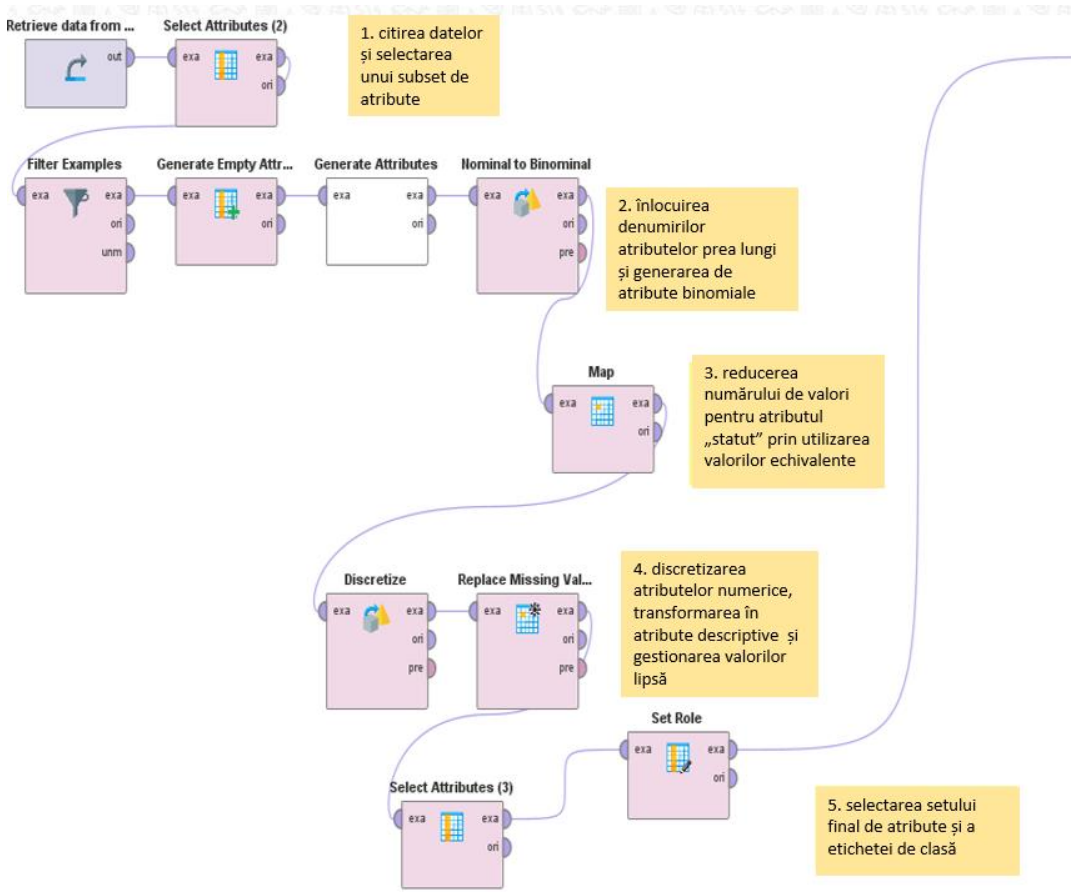


Fig. 6.1 Procesul general de pregătire a datelor structurate pentru modelare

Am ales pentru modelare clasificarea și am testat următoarele tehnici: arbori de decizie (DT - Decision Tree), Support Vector Machine (SVM), KNN, Naïve Bayes, Random Forest (RF) și Deep Learning (DL). Procesul proiectat în *RapidMiner* pentru modelare este prezentat în Fig. 6.2.

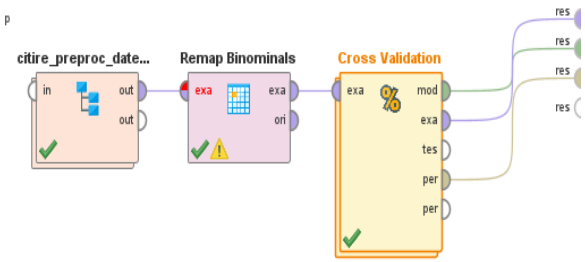


Fig. 6.2 Proiectarea procesului în RapidMiner pentru modelare

Întreaga suită de operații de preprocesare din Fig. 6.1 a fost salvată ca subproces (citire_preproc_date_str). Deoarece majoritatea algoritmilor folosiți lucrează cu clase binare am transformat valorile claselor in binomial. Mai departe, subprocesul *Cross Validation* permite atât antrenarea modelelor cât și testarea acestora și calculul performanțelor.

O captură de ecran pentru conținutul acestui subproces este prezentat în Fig. 6.3. Trebuie specificat că am folosit, în această figură, pentru exemplificare operatorul care implementează algoritmul Naïve Bayes, dar procesul se va repeta pentru toate tehnicile menționate anterior.

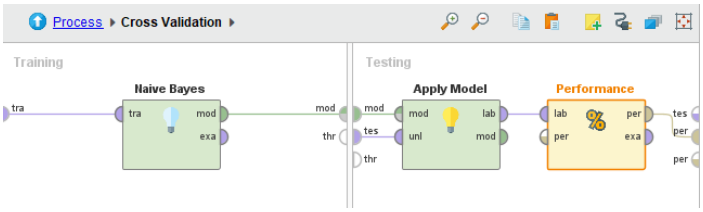


Fig. 6.3 Subproces (citire_preproc_date_str)

6.1.1.1 Modelarea datelor structurate din setul C1

În prima fază, am aplicat procesul de explorare a datelor prezentat în Fig. 6.3 asupra setului de date structurate obținute din chestionarele aplicate profesorilor folosind, pe rând, ca tehnici de modelare: *arbori de decizie* (*DT - Decision Tree), Random Forest (RF), Naïve Bayes, Support Vector machine (SVM), Deep Learning, (DL)* și *KNN*.

A fost folosită o tehnică de inducție a arborilor de decizie cu alegerea atributelor din nodurile de testare pe baza câștigului de informație. Aceasta permite ca, pe baza alegerii pragului minim pentru câștigul de informație să se construiască arbori diferiți, fiecare cu propriile performanțe. S-au obținut performanțele din Tabel 6.1.

Tabel 6.1 Performanțele obținute cu *Decision Tree*, în funcție de valoarea câștigului de informație

Tehnica folosită	Min. gain	Acuratețe [%]	Precizie [%]	Recall [%]	Scor F1 [%]
Decision Tree	0.01	65.48	65.42	56.14	60.04
	0.02	69.14	69.24	61.46	64.46
	0.03	70.81	**69.62**	66.57	67.79
	0.04	**71.13**	69.50	**68.57**	**68.94**
	0.05	**71.13**	69.50	**68.57**	**68.94**

Se poate observa că cea mai bună acuratețe este asociată cu min_gain=0.4 și rămâne constantă și pentru min_gain=0.5. Cu toate acestea, arborii generați nu sunt de cea mai bună calitate (au putere mică de discriminare și posibilități reduse de generare de reguli, sau chiar nu are loc divizarea). În acest context, am ales ca model de considerat mai departe, arborele care corespunde valorii min_gain=0.3, care, deși are acuratețe un pic mai mică, are precizia cea mai mare.

Pentru modelarea cu *Random Forest*, am făcut cercetări experimentale și a rezultat că performanțele cresc daca se optează pentru construcția arborilor prin inducție urmată de tăiere Valoarea optimă a parametrului min_gain este de asemenea 0.3, așa cum se prezintă în Tabel 6.2.

Tabel 6.2 Performanțe Random Forest

Tehnica folosită	Min. gain	Acuratețe [%]	Precizie [%]	Recall [%]	Scor F1 [%]
Random forest	-	66.94	65.73	63.23	64.28
Random forest (cu tăiere)	0.01	70.60	70.61	64.79	67.33
	0.02	71.12	70.81	65.67	67.96
	0.03	**71.54**	**71.30**	**66.34**	**68.51**
	0.04	70.29	70.25	64.33	66.96
	0.05	70.29	70.41	64.11	66.88

S-a considerat pentru analiza ulterioară modelul construit cu valoarea parametrului *minimal gain* = 0,03.

O sinteză a performanțelor obținute prin construirea de clasificatori cu fiecare din tehnicile considerate este prezentată în Tabel 6.3.

Tabel 6.3 Performanțe obținute cu tehnicile utilizate

Tehnica folosită	Acuratețe [%]	Precizie [%]	Recall [%]	Scor F1 [%]
Decision Tree	71.13	69.50	68.57	68.94
Random Forest (cu tăiere)	**71.54**	**71.30**	66.34	68.51
Naïve Bayes	69.65	67.38	68.57	67.76
SVM	67.99	62.01	**82.82**	**70.82**
Deep Learning	67.78	64.14	72.15	67.73
KNN	68.61	67.69	64.82	65.95

Concluzia care se desprinde din această parte a cercetărilor este aceea că, în cazul setului de date corespunzător răspunsurilor profesorilor cele mai bune valori pentru acuratețe și precizie se obțin prin modelarea cu *Random Forest*, iar valorile maxime pentru recall și scorul F1 sunt obținute pentru clasificatorul *SVM*.

6.1.1.2 Modelarea datelor structurate din setul colectat din răspunsurile elevilor (C2)

După o prealabilă etapă de curățare și etichetare a datelor am realizat procese de simulare celor anterioare pentru setul de date structurate culese din chestionarele adresate elevilor. Rezultate obținute referitor la performanțele modelelor construite prin diferite tehnici sunt prezentate în Tabel 6.4.

Tabel 6.4 Performanțele modelelor construite prin tehnicile utilizate

Tehnica folosită	Acuratețe [%]	Precizie [%]	Recall [%]	Scor F1 [%]
Decision Tree	69.89	**69.68**	**81.50**	**70.81**
Random Forest	70.72	67.67	70.01	68.64
Naïve Bayes	**70.82**	67.28	73.05	69.93
SVM	67.90	69.40	54.70	61.01
Deep Learning	69.63	64.61	78.20	70.47
KNN	67.46	63.61	69.25	66.25

Deoarece precizia, recall-ul și scorul F1 au fost maxime pentru clasificatorul bazat pe *arbori de decizie*, s-a considerat acesta ca variantă de lucru ulterioară.

6.1.1.3 Modelarea datelor structurate din setul colectat din răspunsurile părinților (C3)

Datele colectate din chestionarele completate de părinți au trecut prin procese de pregătire similare celor anterioare. Modelarea a respectat, de asemenea, metodologia de la setul colectat din chestionarele profesorilor.

Pentru clasificatorii de tip *Decision Tree* am comparat performanțele obținute cu valori diferite ale parametrului *minimum gain*. Cea mai bună acuratețe (73,22%) a fost obținută pentru min. gain = 0.04.

Similar, pentru Random Forest am implementat un proces cu arbori fără tăiere și unul cu arbori pentru care se practică procedura de tăiere. Cea mai bună acuratețe (75,13%) a fost obținută pentru valoarea 0.01 a parametrului *minimum gain*.

O evaluare cumulativă a performanțelor pentru toate tehnicile de clasificare considerate, în care pentru *Decision Tree* și *Random Forest* am considerat cele mai favorabile rezultate este prezentată în Tabel 6.5.

Tabel 6.5 Performanțele cumulate ale clasificatorilor utilizați

Tehnica folosită	Acuratețe [%]	Precizie [%]	Recall [%]	Scor F1 [%]
Decision Tree	73.22	76.90	88.75	82.33
Random Forest (cu tăiere)	**75.13**	76.65	**93.09**	**84.02**
Naïve Bayes	71.94	78.46	82.76	80.49
SVM	67.22	72.90	88.57	77.96
Deep Learning	75.00	76.87	92.57	83.92
KNN	70.66	**78.49**	80.25	79.29

Rezultatele prezentate arată că și pentru acest set cel mai potrivit algoritm de clasificare este *Random Forest*.

Un studiu comparativ, în ceea ce privește alegerea clasificatorilor pentru cele trei seturi de date este prezentat în Tabel 6.8.

Tabel 6.6 Studiu comparativ privind alegerea clasificatorilor

Setul de date	Clasificatorul ales	Acuratețe [%]	Precizie [%]	Recall [%]	Scor F1 [%]
C1	RF cu tăiere și min_gain=0,03	71.54	71.30	66.34	68.51
C2	DT cu min_gain=0.03	69.89	62.68	81.50	70.81
C3	*Random forest (cu tăiere)*	75.13	76.65	93.09	84.02

Este de menționat faptul că pentru toate cele trei seturi de date, cele mai performante tehnici de clasificare sunt bazate pe *arbori*.

6.1.2 Cercetări privind posibilitățile de creștere a performanțelor modelării ca efect al reducerii dimensionalității seturilor de date

Este de notat faptul că folosind seturile de date structurate din cele trei chestionare în integralitatea lor, ceea ce permite algoritmilor să elimine caracteristicile care nu corespund criteriilor considerate pentru fiecare caz în parte, performanțele modelelor sunt relativ modeste.

Ca urmare, am procedat la reducerea dimensionalității setului de date, folosind două *wrappere*. Avantajul acestora constă în faptul că selecția caracteristicilor este însoțită de procesul de modelare și de calculul performanțelor, ceea ce ușurează proiectarea și implementarea proceselor. Am folosit doi operatori din *RapidMiner Studio 10.3 – Forward Selection*, respectiv *Backward elimination*. Ca măsuri pentru performanță au fost folosite: *acuratețea, precizia, recall și scorul F1*.

Procesul proiectat în *RapidMiner Studio 10.3* este prezentat în Fig. 6.4. Atât *Forward Selection*, cât și *Backward elimination* sunt, la rândul lor, subprocese care încapsulează, în cazul nostru, câte un operator *Cross Validation*, așa cum se arată în Fig. 6.5 pentru *Forward selection*. La rândul lui, acesta este un subproces care conține operațiile deja prezentate în Fig. 6.3.

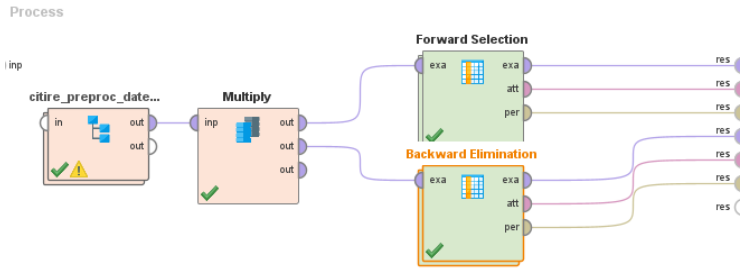


Fig. 6.4 Procesul complet de preprocesare/optimizare a setului de date

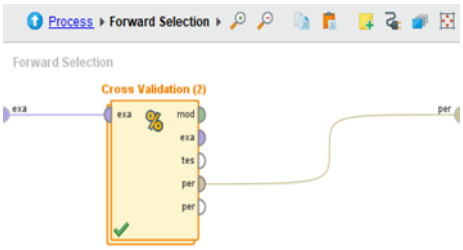


Fig. 6.5 Conținutul subprocesului *Forward selection*

Am folosit un operator de multiplicare cu scopul de a realiza într-un singur proces reducerea dimensionalității prin ambele metode.

6.1.2.1 Modelarea seturilor de date structurate cu dimensionalitate redusă

Am considerat procesul prezentat în Fig. 6.4 pentru cele trei seturi de date C1, C2 și C3.

Performanțele modelelor rezultate în cazul aplicării fiecăreia din cele două wrappere pe fiecare din cele trei seturi de date C1, C2 și C3 sunt:

- Pe setul de date C1:
  a) cu metoda *Forward Selection*, ***SVM*** a obținut cea mai mare acuratețe dintre toate metodele (71,45%) și un scor F1 de 68,99%. Acesta a folosit 4 predictori, un număr relativ mic, ceea ce arată eficiența în selectarea caracteristicilor esențiale;
  b) cu metoda *Backward elimination*, ***Random Forest*** are cea mai bună acuratețe (71,85%), dar un scor F1 mai mic (68,72%), sugerând că modelul poate fi mai sensibil la modul în care sunt selectați predictorii. Astfel, performanța modelului poate varia semnificativ în funcție de predictorii incluși sau excluși din setul de date folosit pentru antrenarea modelului.
- Pe setul de date C2:
  a) cu *Forward Selection*, ***Random Forest*** a obținut cea mai bună acuratețe, 72,23%;
  b) cu *Backward Elimination, Decision Tree* a obținut cele mai bune rezultate în termeni de acuratețe respectiv 71,26% și scor F1, 70.59%.
  Aceste rezultate sugerează că aceste metode sunt bine adaptate pentru problemele de clasificare în educație, menținând un echilibru bun între complexitate și performanță.
- Pe setul de date C3:
  a) cu *Forward Selection*, ***Random Forest*** a obținut cele mai bune rezultate (acuratețe 75,38%, scor F1 de 84,07%, recall de 92,21%), demonstrând o performanță echilibrată și robustă. Cu doar 4 caracteristici selectate, *Random Forest* a reușit să capteze complexitatea datelor și să ofere predicții precise, fiind astfel o alegere bună pentru analiza datelor educaționale, unde acuratețea și capacitatea de generalizare sunt esențiale;
  b) cu *Backward Elimination,* cele mai bune rezultate sunt obținute de ***Deep Learning*** (acuratețe de 76,02%, scor F1 de 84,60%, recall de 93,48%). Aceste performanțe îl fac foarte potrivit pentru aplicarea în domeniul educațional, unde detectarea corectă a cazurilor pozitive este importantă pentru intervenții și decizii informate.
  Aceeași acuratețe, 76,02%, a fost obținută și de ***Random Forest*** care atinge și cel mai mare scor F1 dintre toate modelele (84,65%). Creșterea numărului de caracteristici la 9 a permis modelului să capteze mai bine variabilitatea din date, îmbunătățind atât recall-ul, cât și precizia. Aceasta face din Random Forest un model

bun pentru acest set de date, capabil să gestioneze eficient complexitatea datelor și să ofere predicții precise și robuste.

Sintetic, performanțele modelelor construite pe cele trei seturi de date reduse sunt prezentate în Tabel 6.7.

Tabel 6.7 Performanțele modelelor construite pe cele trei seturi de date reduse

Set de date	Metoda de reducere a dimensionalității	Tehnica de clasificare	Acuratețea [%]	Nr predictori considerați
C1	***Backward elimination***	Random Forest	71.85	17
C2	***Forward selection***	Random Forest	72.23	4
C3	***Backward elimination***	Deep learning	76.02	8
		Random Forest	76.02	9

Valorile obținute pe datele a căror dimensionalitate a fost redusă sunt superioare cu 0.31 – 2.34 procente față de cele obținute pe seturile complete de date.

Rezultatele obținute indică faptul că utilizarea tehnicilor și metodelor de data mining, alături de selecția adecvată a caracteristicilor, poate conduce la dezvoltarea unor modele de clasificare robuste și eficiente în domeniul educațional, contribuind la o mai bună înțelegere și îmbunătățire a procesului de învățare.

În această secțiune am construit modele de clasificare a opiniilor profesorilor, elevilor și părinților pe baza caracteristicilor descriptive polinomiale din seturile de date considerate. Am comparat performanțele modelelor pentru fiecare set de date și am ales acele tehnici considerate a fi cele mai bune.

6.1.3 Pregătirea și modelarea datelor de tip text colectate din chestionarele C1, C2, C3

O parte din întrebările din chestionare au furnizat date de tip text, provenite fie din răspunsuri complet libere, fie din răspunsuri care au presupus alegeri multiple din formulări predefinite.

În cazul răspunsurilor libere, în prima fază a fost necesară o revizuire a ortografiei și rezolvarea situațiilor în care au fost folosiți termeni de jargon sau argou.

Pentru a constitui datele supuse modelării am considerat atributele *avantaje*, *dezavantaje* și *sugestii* pe care le-am concatenat cu scopul obținerii unor documente de consistență sporită.

Limba româna prezintă o serie de particularități care au constituit provocări în procesul de data mining. În primul rând este vorba despre utilizarea diacriticelor, pe care am rezolvat-o alegând ca sistem de codificare ***UTF-8***.

O altă provocare a fost legată de modul în care se pot gestiona negațiile [53] care au efect puternic asupra analizei sentimentelor. Deoarece limba română are particularități multiple și la nivel gramatical, am ales scanarea textelor din documente și înlocuirea negațiilor cu expresii echivalente semantic sau cu antonime. În această etapă, procesul a fost realizat de un colectiv format din două persoane, iar acolo unde au fost găsite diferențe a fost realizat ulterior consensul.

A fost urmat procedeul clasic de transformare a unui text în vector de cuvinte (reprezentate prin valori numerice) utilizând operații de tokenizare, aducere la același tip de caractere (în cazul nostru, la caractere mici), eliminarea cuvintelor de stop și filtrarea token-urilor după lungime și stemming. Pe lângă acești pași care sunt caracteristici preprocesării datelor text, am stabilit atributul etichetă.

Procesul proiectat pentru preprocesarea datelor este prezentat în Fig. 6.6.

Se observă că operatorul *Documents from data* este, de fapt, un subproces al cărui conținut este prezentat în Fig. 6.7.

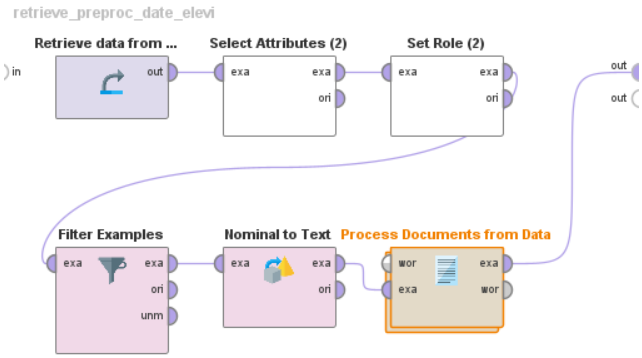


Fig. 6.6 Preprocesarea setului de răspunsuri libere

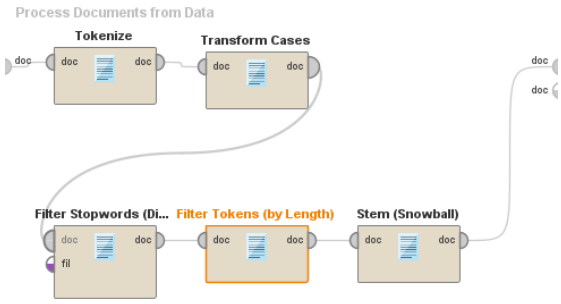


Fig. 6.7 Conținutul subprocesului *Documents from data*

Operațiile de preprocesare au fost aplicate tuturor celor trei seturi de date culese în timpul pandemiei (C1, C2, C3) – care au fost etichetate.

În etapa de modelare am avut ca scop analiza sentimentelor celor trei categorii de persoane implicate în procesul de învățare, pornind de la opiniile lor liber exprimate sub formă de răspunsuri text, folosind tehnicile de data mining potrivite clasificării.

În această cercetare am utilizat tehnicile *Decision Tree, Random Forest, Naïve Bayes, SVM și Deep Learning*.

6.1.3.1 Modelarea setului de date text colectat din răspunsurile profesorilor

Pentru setul de date provenit din răspunsurile profesorilor la chestionarul C1, sinteza performanțelor de clasificare este prezentată în Tabel 6.8.

Tabel 6.8. Performanțe clasificatori pentru setul de date text - profesori (C1)

Tehnica folosită	Acuratețe [%]	Precizie [%]	Recall [%]	Scor F1 [%]
Decision Tree	77,72	72,55	84,84	78,13
Random Forest	81,48	**83,28**	76,85	79,4
Naïve Bayes	54,07	53,08	23,39	32,30
SVM	**81,91**	77,01	**88,20**	**82,12**
Deep learning	81,70	81,08	80,18	80,46

Alegând ca măsură acuratețea, clasificatorul SVM oferă cele mai bune rezultate.

6.1.3.2 Modelarea setului de date text colectat din răspunsurile elevilor

Rezultatele privind performanțele modelelor construite din datele colectate din C2 sunt prezentate în Tabel 6.9.

Tabel 6.9 Performanțe clasificatori pentru setul de date elevi (C2)

Tehnica folosită	Acuratețe [%]	Precizie [%]	Recall [%]	Scor F1 [%]
SVM	81,79	**82,87**	76,54	**79,30**
Naïve Bayes	51,95	48,46	**85,90**	62,26

Decision tree	70,84	68,76	68,31	68,35
Random forest	73,55	75,39	65,03	69,19
Deep learning	**82,18**	80,89	80,75	80,65

Raportat la acuratețe, cea mai bună performanță o obține *Deep learning*.

6.1.3.3 Modelarea setului de date text colectat din răspunsurile părinților

Cercetările asupra modelelor induse prin aplicarea asupra datelor de tip text culese din C3 – chestionarul destinat părinților, au relevat performanțele din Tabel 6.10.

Tabel 6.10 Performanțe clasificatori pentru setul de date părinți (C3)

Tehnica folosită	Acuratețe [%]	Precizie [%]	Recall [%]	Scor F1 [%]
Decision tree	72,83	74,53	93,83	82,94
Random forest	73,73	73,01	**99,64**	84,25
Naïve Bayes	67,87	74,38	80,22	75,64
SVM	81,88	80,34	98,54	88,48
Deep learning	**86,91**	**91,20**	90,22	**90,47**

Similar cazului pentru setul de date aferent elevilor, în cazul clasificatorilor pentru setul de date curent, cea mai bună acuratețe a fost obținută prin *Deep learning*.

6.1.4 Procesarea și modelarea seturilor de date complexe – combinație între date structurate și nestructurate

În această etapă, așa cum a fost menționat în obiective, se realizează prelucrarea combinată a celor două tipuri de date – structurate și nestructurate, care au fost aduse prin preprocesare la formate compatibile.

Procesul de pregătire a datelor pentru modelare este mult mai elaborat (așa cum se prezintă în Fig. 6.8)

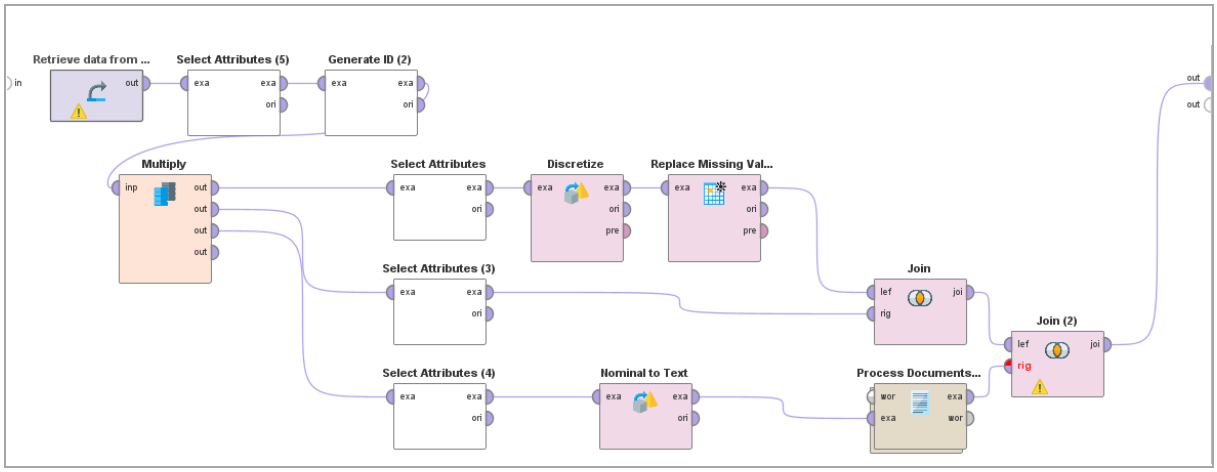


Fig. 6.8 Procesul de pregătire a datelor pentru modelare

Procesul de pregătire a datelor a fost proiectat astfel încât să fie permisă o separare a tipurilor de date în date structurate descriptive, date structurate numerice care necesită discretizare și transformare în date descriptive și date text care necesită propriul mod de

pregătire. La finalul etapelor de pregătire corespunzătoare tuturor acestor categorii se face o joncțiune a rezultatelor astfel încât să se obțină setul complet de atribute ce descriu fiecare caz în parte. Operația de joncțiune necesită existența unei condiții de legătură între date, scop în care am generat, înaintea divizării setului complet de date, un atribut de tip identificator (ID), care a fost preluat în toate cele trei seturi prelucrate individual.

Procesul complet de modelare pentru setul complet de date este prezentat în Fig. 6.9.

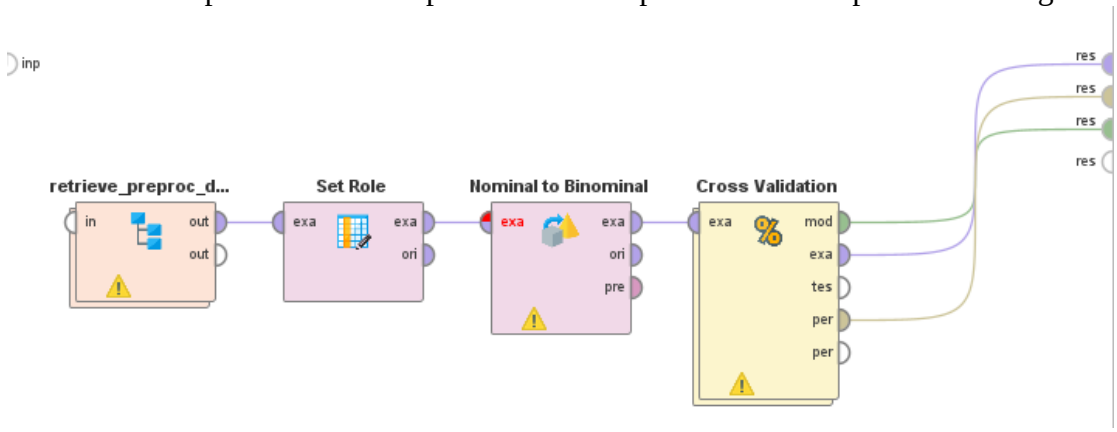


Fig. 6.9 Procesul complet de modelare pentru setul complet de date

În continuare, se consideră toate cele trei seturi C1, C2, C3 pentru care se realizează preprocesarea și construcția de modele prin tehnicile abordate deja: *Decision Tree*, *Random Forest*, *Naïve Bayes*, *KNN*, *SVM* și *Deep Learning*.

6.1.4.1 Modelarea setului complet de date C1

Pentru setul de date provenit din răspunsurile profesorilor la chestionarul C1, pe setul complet de date, sinteza performanțelor de clasificare este prezentată în Tabel 6.11.

Tabel 6.11. Performanțe clasificatori pentru setul de date profesori (C1)

Tehnica folosită	Ac [%]	Prec [%]	Recall [%]	Scor F1 [%]
Decision Tree	77.51	72.95	**83.07**	77.62
Random Forest	**81.27**	**84.59**	74.15	**78.75**
KNN	67.25	67.02	60.35	63.35
Naïve Bayes	52.81	50.13	30.09	37.39
SVM	75.83	71.89	80.41	75.71
Deep Learning	71.33	73	62.83	67.1

Random Forest a obținut cele mai bune rezultate generale dintre toate modelele testate, cu o acuratețe de 81.27% și un scor F1 de 78.75%. Precizia de 84.59% sugerează că acest model este foarte eficient în evitarea cazurilor fals pozitive, însă are un recall mai scăzut (74.15%), ceea ce înseamnă că unele cazuri pozitive ar putea fi ratate. În general, Random Forest s-a dovedit a fi cel mai performant model pentru acest set de date, fiind capabil să generalizeze bine pe date noi.

6.1.4.2 Modelarea setului complet de date C2

Pentru setul de date provenit din răspunsurile elevilor la chestionarul C2, pe setul complet de date, sinteza performanțelor de clasificare este prezentată în Tabel 6.12.

Tabel 6.12 Performanțe clasificatori pentru setul de date elevi (C2)

Tehnica folosită	Acuratețe [%]	Precizie [%]	Recall [%]	Scor F1 [%]
Decision Tree	74.57	76.15	76.02	75.97
Random Forest	74.54	76.17	76.89	76.44
KNN	70.41	70.97	75.99	73.28
Naïve Bayes	59.01	57.96	86.84	69.45
SVM	**80.42**	**78.57**	**87.69**	**82.76**
Deep Learning	77.02	77.62	80.82	79.01

SVM a obținut cele mai bune rezultate dintre toate modelele testate, cu o acuratețe de 80.42% și un scor F1 de 82.76%. Modelul are un recall foarte ridicat (87.69%), ceea ce înseamnă că detectează aproape toate cazurile pozitive, și o precizie solidă (78.57%), indicând un număr relativ mic de fals pozitive. Aceste performanțe fac din SVM cel mai eficient model pentru acest set de date, fiind ideal pentru situații în care atât acuratețea, cât și detectarea cazurilor pozitive sunt esențiale.

6.1.4.3 Modelarea setului complet de C3

Pentru setul de date provenit din răspunsurile părinților la chestionarul C3, pe setul complet de date, sinteza performanțelor de clasificare este prezentată în Tabel 6.13.

Tabel 6.13. Performanțe clasificatori pentru setul de date părinți (C3)

Tehnica folosită	Acuratețe [%]	Precizie [%]	Recall [%]	Scor F1 [%]
Decision Tree	74.11	74.36	96.75	84.02
Random Forest	74.24	73.41	99.64	84.52
KNN	72.83	78.89	83.87	81.27
Naïve Bayes	38.41	69.21	22.64	33.83
SVM	79.95	79.80	96.01	87.12
Deep Learning	**82.26**	85.03	91.12	87.87

Deep Learning a obținut cele mai bune rezultate dintre toate modelele testate, cu o acuratețe de 82.26% și un scor F1 de 87.87%. Modelul are o precizie foarte ridicată (85.03%) și un recall bun (91.12%), ceea ce indică un echilibru excelent între detectarea corectă a cazurilor pozitive și evitarea fals pozitive. Aceste rezultate sugerează că Deep Learning este cel mai puternic model pentru acest set de date, fiind capabil să captureze complexitatea datelor și să facă predicții precise.

6.1.5 Discuții

Analiza rezultatelor obținute pornește de la cele două întrebări enunțate în debutul cercetării, și anume:

I1: *Dată fiind structura complexă a setului de date care cuprinde atât caracteristici structurate cât și nestructurate, care dintre aceste tipuri oferă modele cu performanțe superioare?*

I2: *Ce tehnici de data mining sunt mai eficiente?*

Răspunsul la întrebarea ***I1*** presupune compararea performanțelor obținute în cazul celor trei seturi de date pe caracteristicile nominale, text și combinate. Se va lua în considerare acuratețea modelelor antrenate. Acestea sunt prezentate în Tabel 6.14.

Pentru a răspunde la întrebarea ***I2***, se va evalua eficiența diferitelor tehnici de data mining aplicate pe seturile de date analizate și se vor identifica metodele care oferă cele mai bune rezultate în contextul datelor disponibile.

Tabel 6.14 Performanțe obținute pe seturile de date C1, C2, C3 pentru date structurate, text și pe set complet de date

Set de date	Tip caracteristici considerate pentru modelare	Tehnica de clasificare utilizată	Acuratețe model
C1	*Date structurate*	Random forest (cu tăiere)	71.54
	Date text	**SVM**	**81.91**
	Set complet de date	Random Forest	81.27
C2	*Date structurate*	Naïve Bayes	70.82
	Date text	**Deep learning**	**82.18**
	Set complet de date	SVM	80.42
C3	*Date structurate*	Decision Tree (min_gain=0.04)	73.22
	Date text	**Deep learning**	**86.91**
	Set complet de date	Deep learning	82.26

Diferențele în acuratețe reflectă adaptabilitatea și puterea de predicție a fiecărei tehnici în funcție de natura datelor și a caracteristicilor considerate pentru modelare.

Deep Learning are o performanță ridicată, în special pentru seturile de date text (C2 și C3) și complete (C3). Acest lucru se datorează capacității tehnicii *Deep Learning* de a extrage și a învăța caracteristici complexe din date neorganizate.

SVM a demonstrat, de asemenea, o performanță consistentă, în special pe date text (pentru C1) și complete (C2).

Acuratețea inferioară obținută de *Random Forest, Decision Tree și Naïve Bayes* pe date structurate poate fi atribuită limitărilor acestor tehnici în captarea complexității și a contextului oferit de datele text sau de seturile de date mixte.

Datele text și seturile de date complete permit modelelor să încorporeze mai multe informații contextuale, ceea ce duce la o acuratețe mai mare în comparație cu datele strict structurate.

În concluzie, tehnica ***Deep Learning*** **este mai eficientă pentru seturile de date complexe** care includ date text și structurate, datorită capacității acestei tehnici de a învăța reprezentări bogate ale datelor. ***SVM*** **a fost, de asemenea, o alegere bună pentru datele text**, indicând flexibilitatea și eficiența sa în acest tip de sarcini de clasificare. În schimb, tehnicile ***Naïve Bayes și Random Forest*** **au avut performanțe mai bune pe date structurate**, dar au fost mai puțin eficiente pe seturi de date mixte sau text.

6.1.6 Contribuții

- Toate procesele au fost proiectate și implementate pe seturi de date proprii, provenite din răspunsurile profesorilor, elevilor și părinților la un set de trei chestionare distribuite în perioada pandemiei, care au vizat identificarea opiniilor referitoare la trecerea bruscă la învățământul online.
- Etichetarea seturilor de date a fost realizată printr-un proces iterativ de către un colectiv format din 3 specialiști: 1 profesor universitar - doctor în calculatoare și tehnologia informației; 1 profesor psihopedagog - doctor în științele educației și 1 doctorand în

domeniul calculatoare și tehnologia informației. Acest proces a implicat mai multe etape: *selecția inițială a etichetelor urmată de multiple etape de revizuire și validare pentru a minimiza subiectivitatea și erorile*. Procesul iterativ de etichetare, important pentru asigurarea calității și acurateței modelului, a contribuit la crearea unui set de date etichetat de calitate, care a fost fundamental pentru antrenarea eficientă a modelelor de învățare automată. Etichetele finale au fost validate prin comparație cu un set de referință. Această validare a confirmat că setul de date a fost bine etichetat și gata pentru utilizarea în antrenarea modelelor de învățare automată. Acest proces a implicat mai multe etape succesive de evaluare și corectare a etichetelor, având ca scop îmbunătățirea continuă a datelor și, implicit, a performanței modelelor de clasificare.

- Utilizarea tehnicilor de text mining pe documente în limba română, deoarece păstrarea limbii în care au fost culese datele pentru crearea modelelor asigură o reprezentare mai apropiată de realitate decât un text tradus. Deoarece în cercetările efectuate nu am găsit un corpus în limba romană care să ne permită antrenarea modelelor de clasificare am realizat atât antrenarea cât și testarea și validarea modelelor pe seturile de date culese. De asemenea, nu am identificat lucrări care să abordeze text mining pe domeniul educației în limba română.
- Realizarea unui nou dicționar de cuvinte de stop în limba română, conform necesităților textului de analizat.
- Cercetări empirice privind acuratețea modelelor construite precum și identificarea și analiza factorilor de influență în calitate de predictori ai modelului de clasificare.

6.2 Identificarea factorilor care influențează alegerea de către profesorii din mediul preuniversitar românesc a formării IT pentru a evita problemele ridicate de educația online

Această secțiune își propune să prezinte un studiu experimental privind modelarea atitudinii profesorilor față de participarea la cursuri de dezvoltare profesională în domeniul IT pentru a oferi lecții online de calitate. Este utilizat setul de date C1 colectate printr-un chestionar la care au răspuns 956 de cadre didactice.

Obiectivul cercetării este ca, după construirea celui mai performant model, să fie analizate caracteristicile care sunt semnificative în construcția sa. Aceste caracteristici sunt factori de influență în alegerile profesorilor.

Am formulat, spre rezolvare, următoarele întrebări de cercetare:

IC1: *Care este tehnica de modelare potrivită pentru a găsi acei factori care influențează decizia profesorilor de a urma cursuri de perfecționare în IT?*

IC2: *Care sunt factorii de influență ai acestei decizii?*

Cu intenția de a construi modele care ar putea dezvălui motivele care stau la baza opțiunii acestora, motive ce pot fi analizate ulterior ca factori care pot fi folosiți pentru a influența pozitiv sistemul, am considerat răspunsurile la întrebarea numărul 20 *(”Participarea la cursuri de formare continuă în domeniul IT v-ar ajuta să gestionați mai bine resursele software și hardware necesare școlii digitale?”)* ca etichetă de clasă.

Prelucrarea datelor

Cu scopul de a pune datele într-o formă adecvată pentru modelare, datele au fost preprocesate.

În acest scop, au fost efectuate următoarele operațiuni:

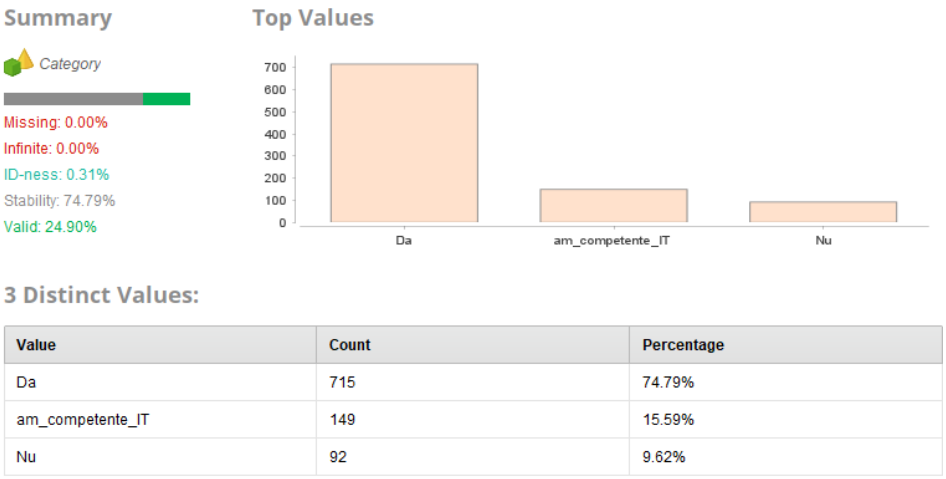
- am verificat dacă toate câmpurile sunt completate;

- am făcut modificări în câmpurile care conțin date redundante (de exemplu, am clasificat cadrele didactice cu gradul I; am verificat dacă acestea aveau și studii doctorale pentru a le integra în categoria cu cel mai mare grad didactic);pentru întrebările la care s-a răspuns pe o scală de la 1 la 10, am înlocuit numerele cu calificative, astfel: 10 a devenit E (excelent), 8-9 - FB (foarte bine), 7 - B (bine), 1-6 - S (suficient);
- pentru întrebările cu valori ale răspunsului între 1 și 5, am înlocuit cifrele cu calificative după cum urmează: 5 a devenit E (excelent), 4 - 3 FB (foarte bun), 2 - B (bun), 1 - S (suficient);
- la întrebarea ”Poate școala online să înlocuiască școala tradițională?” pentru care răspunsurile stabilite au fost ”da”, ”nu” și ”nu știu”, am înlocuit ”nu știu” cu ”mă abțin”;
- am simplificat formularea întrebărilor din chestionar (pe care le-am folosit în procesare ca etichete sau atribute) prin abrevieri sugestive, de exemplu, întrebarea ”Apreciați competențele digitale pe care le aveți” a fost înlocuită cu ”Level_comp_IT”.

În plus, prin utilizarea operatorilor corespunzători furnizați de mediul *RapidMiner Studio 10*, au fost eliminate caracteristicile puternic corelate, iar atributele cu foarte multe valori distincte (care nu au o pondere semnificativă în inducerea modelelor de clasificare), precum și atributele cu o pondere crescută a aceleiași valori au fost, de asemenea, eliminate.

După cum se prezintă în Fig. 6.10, valorile claselor sunt puternic dezechilibrate.

Deoarece acest dezechilibru poate afecta modelul, am propus două metode de echilibrare a acestora: prin *subeșantionare* pentru cazurile cu etichetele „Da” (“Yes”) și „am_competente_IT” („I am_competent_IT”) și prin *supraeșantionare* pentru cele cu etichetele “Nu” („No”) și „am_competente_IT” („I am_competent_IT”).



Value	Count	Percentage
Da	715	74.79%
am_competente_IT	149	15.59%
Nu	92	9.62%

Fig. 6.10 Distribuția valorilor claselor din setul de date original

O captură de ecran cu procesul de subeșantionare este prezentat în Fig. 6.11. Am verificat datele privind echilibrul parametrilor prin intermediul operatorului *Sample* și am redus numărul de cazuri pentru fiecare dintre cele trei clase vizate la 92.

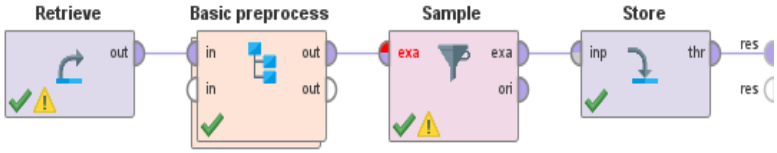
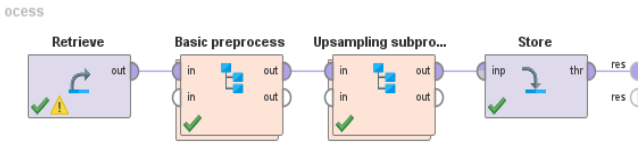
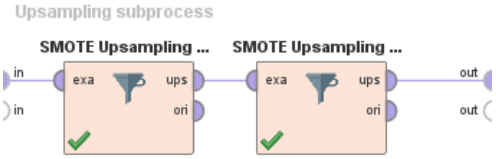


Fig. 6.11 Echilibrarea datelor prin subeșantionare

Pentru a echilibra datele prin supraeșantionare, am utilizat operatorul *SMOTE Upsampling*, care utilizează tehnica de supraeșantionare minoritară sintetică [54]. Procesul este prezentat în Fig. 6.12.



a)



b)

Fig. 6.12 Echilibrarea datelor prin supraeșantionare: a) întregul proces, b) componentele subprocesului de supraeșantionare

După rularea celor două procese, au rezultat două seturi de date echilibrate, unul cu 276 de cazuri, celălalt cu 2145 de exemple.

Modelarea datelor

Am folosit clasificarea și, în prima etapă, am luat în considerare următoarele clasificatoare: *Naïve Bayes, Decision Tree, Random Forest și SVM*.

Utilizând setul de date original am obținut valorile de performanță din Tabel 6.15.

Tabel 6.15 Performanțele procesului pe setul de date original

Tehnica folosită	Acuratețe [%]	Timp de execuție [s]
Naïve Bayes	73.6%	10
Decision Tree	79.5%	9
Random Forest	**79.9%**	16
SVM	78.8%	37

Cea mai bună performanță pe setul de date original, 79,9%, o obține *Random Forest*.

Rulând în aceleași condiții seturile obținute prin cele două metode de echilibrare, am obținut rezultatele prezentate în Tabel 6.16 și Tabel 6.17.

Tabel 6.16 Performanțele proceselor pe setul de date supraeșantionate

Tehnica folosită	Acuratețe [%]	Timp de execuție [s]
Naïve Bayes	57.7%	7
Decision Tree	62.1%	8
Random Forest	**74.7%**	2min 16 sec
SVM	72.6%	1 min 45 sec

Cea mai bună performanță pe setul de date supraeșantionate, 74,7%, o obține *Random Forest*.

Tabel 6.17 Performanțele proceselor pe setul de date subeșantionate

Tehnica folosită	Acuratețe [%]	Timp de execuție [s]
Naïve Bayes	48.7%	10
Decision Tree	20.5%	9
Random Forest	34.9%	16
SVM	**50.6%**	37

Cea mai bună performanță pe setul de date subeșantionate, 50,6%, o obține SVM.

Atât pentru setul de date original, cât și pentru setul de date supraeșantionat, performanța este considerabil mai bună decât cea obținută pe setul de date subeșantionat și echilibrat. Cu toate acestea, cele mai bune rezultate nu depășesc pragul de 79,9% pentru acuratețe și, paradoxal, cele mai bune valori sunt obținute pe setul de date dezechilibrat. De asemenea, analizând datele din tabelele de mai sus, am constatat că cele mai bune valori au fost obținute pentru modelarea *Random Forest*. Ca rezultat, am considerat în continuare posibilitățile de a crește eficiența acestor modele, așa că am evaluat diferite combinații de valori ale parametrilor, dar performanța nu s-a schimbat semnificativ.

Conform celui mai utilizat model pentru procesele de data mining, - CRISP-DM [55], aceste procese sunt puternic interactive și iterative. Deoarece în primele două iterații rezultatele pot fi considerate satisfăcătoare, dar nu foarte bune, am continuat studiul experimental într-o nouă direcție, luând în considerare doar cazurile seturilor de date originale și supraeșantionate și tehnica de data mining *Random Forest*. Am abordat un cadru care utilizează două metode de reducere a dimensionalității. Mai întâi, am efectuat o selecție a caracteristicilor utilizând metode *wrapper* care evaluează toate combinațiile posibile de caracteristici și o selectează pe cea care produce cele mai bune rezultate pentru algoritmul de învățare automată vizat.

Am considerat cele două scenarii posibile: selecție înainte (*forward selection*) și eliminare inversă (*backward elimination*). Rezultatele sunt prezentate în Tabel 6.18.

Prin această operațiune am avut ca scop utilizarea celei mai bune combinații de caracteristici, cele care participă cu adevărat la construirea modelului.

Tabel 6.18 Rezultatele selecției de caracteristici utilizând *wrappers*

Set de date	Metoda de selecție	Nr. predictori selectați	Acuratețe
Original	Forward selection	**4**	**88.05 %**
	Backward elimination	23	88.05 %
Upsampled	Forward selection	10	87.20%
	Backward elimination	**14**	**87.69%**

Metoda *Forward Selection* pe setul de date original a selectat doar 4 predictori, ceea ce a dus la obținerea unei acurateți ridicate de 88.05%. Aceasta sugerează că un subset mic și bine ales de caracteristici poate fi suficient pentru a construi un model eficient. Acest rezultat este foarte valoros, indicând că complexitatea modelului poate fi redusă semnificativ fără a compromite performanța.

În contrast, *Backward Elimination* a selectat un număr mult mai mare de predictori (23), dar a obținut exact aceeași acuratețe de 88.05% ca și Forward Selection. Acest lucru poate sugera că mulți dintre predictorii adăugați de Backward Elimination nu aduc un beneficiu

semnificativ în performanța modelului, ceea ce ridică întrebări despre eficiența și necesitatea lor.

Pe setul de date upsampled, Forward Selection a selectat 10 predictori și a obținut o acuratețe ușor mai mică de 87.20%. Aceasta indică faptul că, atunci când datele sunt echilibrate (prin upsampling), modelul poate avea nevoie de un număr mai mare de predictori pentru a menține o performanță ridicată. Totuși, scăderea în acuratețe comparativ cu setul original sugerează că upsampling-ul poate introduce zgomot sau variații care complică clasificarea.

În cazul Backward Elimination pe setul upsampled, au fost selectați 14 predictori, iar acuratețea obținută a fost de 87.69%, ușor mai mare decât cea obținută cu Forward Selection. Aceasta sugerează că metoda Backward Elimination este mai robustă în cazul datelor upsampled, reușind să identifice un set de predictori care echilibrează mai bine modelul, deși acuratețea rămâne totuși ușor mai mică decât cea obținută pe setul original.

În al doilea rând, cu intenția de a identifica cei mai buni predictori, am utilizat metoda eliminării atributelor inutile, adică acele atribute care depășesc pragul de inutilitate stabilit prin parametri adecvați pentru fiecare tip de date. Rezultatele obținute sunt prezentate în Tabelul 6.19.

Tabel 6.19 Rezultatele eliminării caracteristicilor inutile

Set de date	Prag inutilitate	Nr. predictori selectați	Acuratețe
Original	0.1	22	88.11%
	0.2	21	88.11%
	0.3	**17**	**88.11%**
	0.4	15	87.86%
	0.5	12	87.74%
Upsampled	0.1	16	85.8%
	0.2	15	85.8%
	0.3	15	85.8%
	0.4	13	80.14%
	0.5	13	80.14%

În ambele seturi de date, se observă că un prag de inutilitate de 0.3 oferă o combinație optimă între reducerea complexității modelului și menținerea unei acurateți ridicate.

Ultimul pas pe care l-am considerat ca o posibilitate de a crește calitatea modelului este utilizarea operatorului de optimizare automată a parametrilor procesului de data mining. Acest proces de optimizare se bazează pe seturile de date (original și supraeșantionat) care, în pasul anterior, au oferit cele mai bune valori pentru acuratețe. Procesul de optimizare utilizează o execuție iterativă a procesului de modelare folosind toate combinațiile de parametri.

Pe scurt, rezultatele obținute, constând în parametrii optimi, atât pentru seturile de date originale, cât și pentru cele echilibrate, sunt prezentate în Tabelul 6.20.

Tabel 6.20 Rezultatele optimizării automate a parametrilor

Data set	No. of. trees	Splitting criterion	Pre-prunning	Pru-nning	Acuratețe
Original	**21**	**Gain_ratio**	**T**	**F**	**89.4%**
Upsampled	70	Gain_ratio	F	F	86.9%

Pentru setul de date upsampled, modelul a utilizat un număr mai mare de arbori (70), dar nu a aplicat nici pre-tăierea, nici post-tăierea. Acuratețea obținută a fost de 86.9%, ceea ce este ușor mai mică decât acuratețea obținută pe setul de date original. Acest rezultat poate indica faptul că, în cazul datelor upsampled, complexitatea modelului a crescut semnificativ (dat fiind numărul mare de arbori), dar aceasta nu s-a tradus neapărat într-o îmbunătățire a performanței. Faptul că pre-tăierea și post-tăierea au fost dezactivate ar putea sugera că modelul a devenit mai susceptibil la overfitting, mai ales în cazul unui număr mare de arbori.

Inițial, am luat în considerare patru tehnici de clasificare, după care am rafinat procesul de optimizare pentru Random Forest, care a dovedit din primele experimente cea mai bună performanță.

Cercetarea a produs un rezultat surprinzător. În toate scenariile studiate, performanța modelelor construite pe setul de date original a fost superioară celor construite pe seturile de date echilibrate.

Dacă modelele construite pe setul subeșantionat aveau doar 276 de cazuri de antrenament, ceea ce este un număr foarte mic și, prin urmare, rezultatele sunt explicabile, același lucru nu poate fi spus despre cele bazate pe setul supraeșantionat.

În acest caz, motivul ar putea fi legat de metoda utilizată pentru supraeșantionare. Aceasta implică injectarea în setul original a datelor generate sintetic corespunzătoare claselor minoritare. Deși această procedură împiedică modelul să fie părtinitor față de clasa majoritară, în cazul nostru aceasta a dus la o scădere a acurateței.

O altă observație se referă la faptul că o îmbunătățire semnificativă a acurateței modelului, de aproximativ 10%, a fost obținută prin selecția caracteristicilor. Aceasta implică reducerea setului inițial la o cardinalitate minimă. Acest nou set de date trebuie să conțină cele mai relevante atribute pentru scopul modelării. În cele din urmă, optimizarea parametrilor procesului a dus, de asemenea, la o creștere a performanței, dar aceasta a fost la un nivel de aproximativ 1%.

Cercetarea experimentală efectuată a permis găsirea celui mai bun model care să descrie această atitudine și să descopere și să cuantifice factorii care o determină.

Deoarece cel mai bun model constă într-o combinație de 21 de arbori de clasificare, majoritatea având cel puțin 6 niveluri și un număr semnificativ de frunze, pentru a explica predicțiile am considerat ponderile predictorilor în dezvoltarea modelului. Acestea sunt prezentate în Fig. 6.13.

Principalii factori care influențează dorința profesorilor de a-și îmbunătăți predarea online prin îmbunătățirea competențelor IT sunt: „Necesitatea unui cadru legal reglementat pentru învățământul online”, „Mediu”, „Nivel”, „Online_vs_tradițional” și „Nivel_comp_it”.

Dezvoltarea unui cadru legal solid pentru programele de formare IT și educația online în România are implicații politice semnificative, care ar putea influența sistemul educațional și piața muncii pe termen lung. Crearea unui cadru legal pentru educația online și programele IT ar trebui să asigure acces egal pentru profesori la resurse educaționale, indiferent de locația geografică sau statutul socio-economic. Aceasta implică investiții guvernamentale în infrastructura digitală, în special în zonele rurale sau defavorizate, pentru a reduce disparitățile în accesul la educație. Această inițiativă necesită o colaborare strânsă între diverse agenții guvernamentale, sectorul privat și comunitățile academice, subliniind importanța unei abordări holistice și bine coordonate.

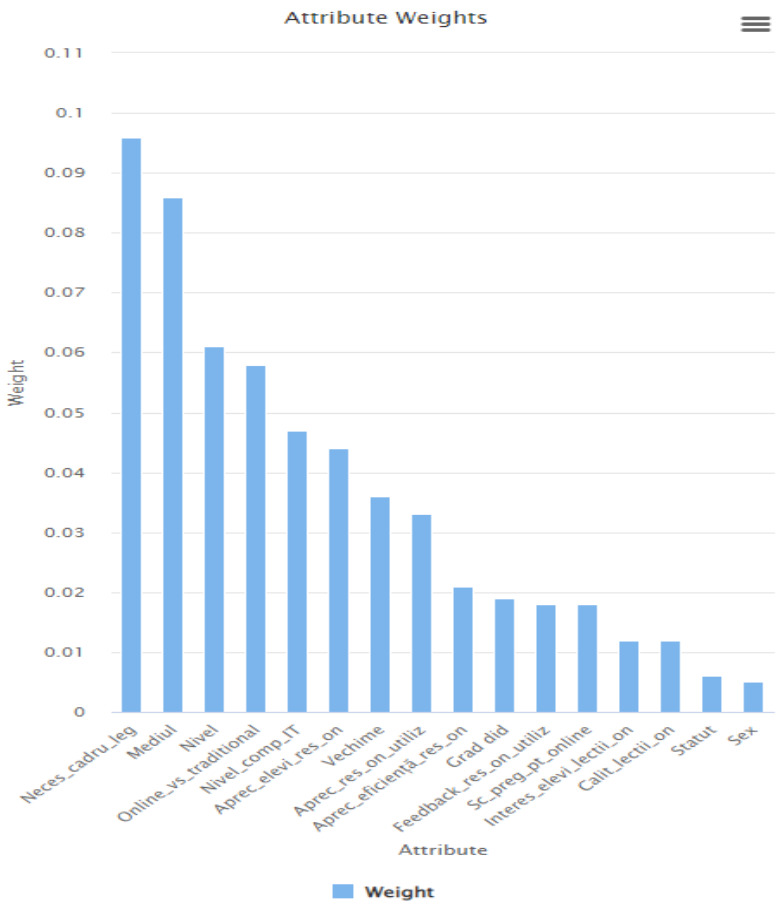


Fig. 6.13 Ponderile predictorilor în cel mai bun model de clasificare

O altă direcție în care sunt posibile intervenții se referă la limitarea diferențelor privind ***mediul*** în care se desfășoară procesul educațional. Discrepanțele dintre profesorii români care predau în mediul urban și cei din mediul rural constă în diferențele semnificative în accesul la resurse și infrastructură tehnologică.

Un alt predictor cu pondere mare este ***nivelul*** la care predau cadrele didactice. Motivul principal al diferențelor dintre profesorii români care predau la nivel preșcolar, primar, gimnazial și liceal constă în variațiile semnificative ale pregătirii profesionale, resurselor disponibile și infrastructurii tehnologice la fiecare nivel educațional.

Preferințele profesorilor români ***pentru învățământul online sau cel tradițional*** variază semnificativ. Cei mai mulți dintre aceștia apreciază interacțiunea directă și adaptabilitatea imediată oferite de predarea față în față, în timp ce alții, mai puțini, valorizează flexibilitatea, personalizarea procesului de învățare și accesul la resurse digitale diverse caracteristice învățământului online.

Diferențele în ***nivelurile de competențe IT ale profesorilor*** influențează în mod direct eficiența și succesul integrării tehnologiilor moderne în educație, subliniind necesitatea unor strategii educaționale și politici de formare care să abordeze aceste disparități și să sprijine dezvoltarea competențelor digitale pentru toți profesorii. Profesorii care au acces regulat la programe de formare pentru integrarea resurselor IT în activitatea didactică sunt mai bine pregătiți pentru a implementa și utiliza instrumentele digitale în educație, spre deosebire de cei care nu aleg să se formeze sau nu beneficiază de oportunități în acest sens.

În acest context, propun următoarele recomandări – corelate cu factorii care influențează percepția negativă a profesorilor privind învățământul online (Tabel 6.21).

Tabel 6.21 Recomandări corelate cu factorii/predictorii care influențează percepția negativă a profesorilor privind învățământul online

Factori/ predictori	**Recomandări**
Dezvoltarea unui cadru legal solid pentru programele de formare IT și educația online în România	• dezvoltarea unui cadru de formare continuă pentru profesorii din învățământul preuniversitar prin organizarea de cursuri și ateliere specializate pe dezvoltarea competențelor IT și utilizarea tehnologiilor educațional. Acest cadru ar trebui să fie accesibil atât pentru cadrele didactice din mediul urban, cât și pentru cele din mediul rural, utilizând platforme online pentru a asigura acoperirea națională și să includă sesiuni practice și exemple concrete de utilizare a tehnologiilor digitale în predare, adaptate diferitelor discipline și niveluri educaționale;
Mediul în care se desfășoară procesul educațional	• prioritizarea investițiilor în infrastructura tehnologică a școlilor, în special în mediul rural, pentru a reduce discrepanțele între unitățile de învățământ. Asigurarea accesului la internet de mare viteză și dotarea cu echipamente IT moderne (tablete, laptopuri, proiectoare, tablă interactivă etc.);
Nivelul la care predau cadrele didactice	• identificarea nevoilor specifice de formare și adaptarea în consecință a programelor de formare; • Implementarea unor programe de mentorat în care profesorii cu competențe IT avansate să ofere sprijin și îndrumare colegilor lor;
Preferințele profesorilor români pentru învățământul online sau cel tradițional	• dezvoltarea unei platforme naționale de resurse educaționale digitale care să fie accesibile gratuit pentru toți profesorii și elevii. Aceasta ar include lecții interactive, tutoriale video, simulări, aplicații educaționale și alte resurse didactice; • încurajarea colaborării între profesori pentru a crea și împărtăși resurse educaționale digitale de calitate, facilitând astfel accesul la material didactic diversificat și actualizat;
Diferențele în nivelurile de competențe IT ale profesorilor	• implementarea de metodologii de evaluare a competențelor digitale ale profesorilor, pentru a identifica nevoile specifice de formare și pentru a monitoriza progresul și utilizarea rezultatelor evaluărilor pentru a personaliza programele de formare și pentru a oferi suport suplimentar acolo unde este necesar; • promovarea și integrarea practicilor de data mining în educație prin încurajarea utilizării tehnicilor de data mining pentru analiza datelor educaționale, cu scopul de a identifica tendințe, a personaliza procesul de învățare și a îmbunătăți rezultatele educaționale; • organizarea de workshop-uri și seminarii pentru a familiariza profesorii cu conceptele de bază ale data mining și pentru a demonstra aplicabilitatea acestora în context educațional; • colaborarea cu instituții de învățământ superior și centre de cercetare prin stabilirea de parteneriate cu universități și centre de cercetare pentru a dezvolta proiecte comune de cercetare și implementare a tehnologiilor educaționale; • evaluarea impactului tehnologiilor educaționale prin realizarea de studii periodice pentru a evalua impactul utilizării tehnologiilor educaționale asupra performanței academice a elevilor și asupra

	eficienței predării; utilizarea rezultatelor acestor studii pentru a adapta și îmbunătăți continuu strategiile de integrare a tehnologiilor în educație.

Aceste recomandări sunt menite să abordeze și să reducă disparitățile în competențele IT ale profesorilor și să sprijine o integrare eficientă și echitabilă a tehnologiilor moderne în sistemul educațional românesc.

6.2.1 Contribuții

- Cercetare empirică privind măsura în care selecția caracteristicilor și optimizarea automată a parametrilor procesului de modelare au un impact asupra performanței generale a modelului (acuratețe sau timp de execuție), în vederea implementării unui cadru care să ofere un raport acceptabil între performanță și consumul de resurse.
- După o analiză atentă a literaturii de specialitate, am constatat că nu au fost realizate încă studii privind factorii care influențează alegerea cursurilor de formare în domeniul IT de către profesori pentru a preveni eventualele probleme ridicate de necesitatea de a preda online.
- Analiza critică a factorilor de influență si elaborarea de recomandări.

6.3 Cercetări privind evoluția percepției asupra învățământului online

Obiectivele propuse în această secțiune sunt:

O1. *Analiza sentimentelor celor trei categorii de respondenți (profesori, elevi, părinți) privind desfășurarea învățământului online*

O2. *Identificarea factorilor/predictorilor/variabilelor care influențează sentimentele negative.*

Pentru cercetare am utilizat seturile de date C4 și C5. Acestea au fost descrise în Capitolul 5, în amănunt.

6.3.1 Educația on-line după un an de la debutul pandemiei Covid-19

Cu scopul evidențierii tendințelor în evoluția factorilor determinanți ai opiniilor negative privind educația online, la distanță de un an de la debutul pandemiei, în anul 2021, a fost cules un set nou de date pe baza unui chestionar distribuit de asemenea celor trei piloni ai resurselor umane din sistemul educațional preuniversitar românesc: profesori, elevi și părinți. Setul rezultat, C4, a fost descris amănunțit în Capitolul 5 din teză.

Atingerea obiectivelor propuse presupune găsirea celor mai bune modele de analiză a sentimentelor celor trei categorii de respondenți privind desfășurarea orelor online și, apoi, din aceste modele identificarea factorilor care influențează sentimentele negative.

C4 conține atât date structurate cât și date text provenite din răspunsuri libere. Într-un prim pas au fost create modele doar din datele structurate, apoi din întregul set de date.

Pentru preprocesare, un prim pas a fost înlocuirea valorilor colectate din chestionare cu șiruri de caractere simple, dar sugestive, astfel: *Total dezacord* - TD, *Dezacord* - D, *Abținere* - AB, *De acord* - A, *Total de acord* -TA.

Au fost respectate aceleași etape prezentate în Fig. 6.1 și Fig. Fig. 6.2, iar modelarea a fost realizată după procese similare celui descris în Fig. 6.3.

Au fost considerate aceleași tehnici de clasificare abordate pentru seturile de date C1-C3, adică: *Decision Tree, Random Forest, KNN, Naïve Bayes, SVM și Deep Learning*.

Rezultatele privind performanțele obținute sunt prezentate în Tabel 6.22, Tabel 6.23 și Tabel 6.24.

Tabel 6.22 Performanțele clasificatorilor pentru răspunsurile profesorilor din setul de date C4

	Set date structurate				**Set complet de date**			
	Acurat [%]	**Prec [%]**	**Recall [%]**	**scor F [%]**	**Acurat [%]**	**Prec [%]**	**Recall [%]**	**scor F1 [%]**
DT	68.15	60.96	31.28	40.97	83.34	85.78	64.56	73.07
RF	**71.98**	71.31	38.71	49.1	86.87	8.21	75.16	80.16
NB	70.17	58.82	57.68	57.89	70.79	59.88	93.95	71.16
KNN	69	58.63	47.62	52.25	72.93	68.66	46.1	54.78
SVM	64.65	52.75	19.46	27.19	80.89	65.98	97.35	78.59
DL	67.3	54.31	60.38	56.88	**88.75**	84.9	84.04	84.28

Random Forest a performat pe setul complet de date, cu o acuratețe de 86.87% și un scor F de 80.16%. Aceasta indică faptul că modelul beneficiază semnificativ de complexitatea și completitudinea setului complet de date. Pe setul structurat, performanța este mai modestă, 71.98%, dar cea mai bună dintre clasificatorii considerați.

Deep Learning a obținut cele mai bune rezultate pe setul complet de date, cu o acuratețe de 88.75% și un scor F1 de 84.28%. Performanțele sale sunt semnificativ mai bune pe setul complet, ceea ce sugerează că acest model beneficiază cel mai mult de datele bogate și complexe. Pe setul structurat, performanța este semnificativ mai slabă, indicând că modelul necesită o cantitate mare de informații pentru a performa optim.

În concluzie, toate modelele au performat mai bine pe setul complet de date comparativ cu setul structurat, ceea ce indică faptul că structurarea datelor poate duce la pierderea de informații esențiale care afectează capacitatea modelelor de a face predicții precise. Structurarea datelor a avut un impact negativ asupra performanței tuturor clasificatorilor, cel mai vizibil în cazul SVM și Deep Learning. Aceste modele, care sunt capabile să capteze relații non-liniare și complexe, par să aibă nevoie de seturi de date complete și bogate pentru a funcționa la capacitate maximă.

Dacă se are în vedere setul de date colectat din răspunsurile elevilor, același clasificator *Deep Learning* este cel mai performant, cu o acuratețe de 91.54% și un scor F1 de 91.77%. Aceste rezultate confirmă capacitatea modelului de a exploata pe deplin complexitatea și bogăția informațiilor din setul complet de date. Îmbunătățirea semnificativă a performanțelor comparativ cu setul structurat indică faptul că modelul de Deep Learning este foarte eficient în gestionarea datelor complexe (Tabel 6.23).

Naïve Bayes a avut performanțe mixte. Pe setul complet de date, acuratețea a scăzut la 55.3%, dar recall-ul a crescut dramatic la 99.72%, sugerând că modelul a devenit foarte sensibil la detectarea cazurilor pozitive, însă cu costul unui număr mare de fals pozitive, ceea ce a dus la o scădere a preciziei. Acest model pare să aibă dificultăți în a generaliza bine pe date complexe și bogate (Tabel 6.23).

Tabel 6.23 Performanțele clasificatorilor pentru răspunsurile elevilor din setul de date C4

	Set date structurate				Set complet de date			
	Acurat [%]	**Prec [%]**	**Recall [%]**	**scor F [%]**	**Acurat [%]**	**Prec [%]**	**Recall [%]**	**scor F1 [%]**
DT	61.64	64.85	53.19	58.24	80.99	74.77	95.14	83.61
RF	65.06	65.84	64.63	65.16	87.52	86.9	88.77	87.81
NB	**65.32**	65.87	65.64	65.71	55.3	53.13	99.72	69.32
KNN	63	63.53	63.57	63.46	70.64	74.08	64.63	68.97
SVM	61.03	71.26	39.11	50.38	71.51	64.08	99.83	78.04
DL	63.14	61.36	73.68	66.86	**91.54**	90.68	92.96	91.77

Așa cum se prezintă în Tabel 6.24, pentru răspunsurile părinților din setul complet de date C4, *Deep Learning* a obținut cele mai bune rezultate, cu o acuratețe de 93.15% și un scor F1 de 91.31%. Aceste rezultate confirmă capacitatea modelului de a exploata pe deplin complexitatea și bogăția informațiilor din setul complet de date. Îmbunătățirea semnificativă a performanțelor comparativ cu setul structurat indică faptul că modelul de Deep Learning este foarte eficient în gestionarea datelor complexe.

Random Forest a obținut rezultate foarte bune pe setul complet de date, cu o acuratețe de 90.2% și un scor F de 87.02%. Îmbunătățirea semnificativă față de setul structurat (acuratețe de 67.35%) indică faptul că modelul a putut extrage și utiliza informații relevante din setul complet, care nu erau disponibile sau evidente în setul structurat. Modelul a devenit mult mai robust și capabil să identifice corect atât cazurile pozitive, cât și cele negative.

Tabel 6.24 Performanțele clasificatorilor pentru răspunsurile părinților din setul de date C4

	Set date structurate				Set complet de date			
	Acurat [%]	**Prec [%]**	**Recall [%]**	**scor F1 [%]**	**Acurat [%]**	**Prec [%]**	**Recall [%]**	**scor F1 [%]**
DT	66.26	62.95	33.04	43.22	87.66	92.61	74.97	82.36
RF	**67.35**	63.54	39.15	48.29	90.2	89.85	84.42	87.02
NB	67.13	57.26	63.94	60.29	50.48	44.1	99.92	61.18
KNN	64.53	56.32	41.14	47.49	72.25	70.76	49.8	58.31
SVM	60.96	-	-	-	72.09	58.4	100	73.71
DL	64.99	54.27	66.95	59.79	**93.15**	90.53	92.13	91.31

În Tabel 6.25 sunt prezentați factorii care au pondere semnificativă în inducerea modelelor de clasificare pentru setul de date C4.

Tabel 6.25 Predictorii care au pondere importantă în crearea modelelor pentru setul de date C4

Set de date	**Predictori / factori**	**Opțiune respondent**
Profesori - C4	Activitățile online sunt plictisitoare și monotone pentru elevi în lipsa interacțiunii directe cu colegii și profesorii .	Dezacord
	Activitățile online sunt mult mai atractive pentru elevi.	Total de acord

Set de date	Predictori / factori	Opțiune respondent
	Pe ansamblu, cred că în ultimul an s-au făcut progrese foarte mari în creșterea calității și rezultatelor învățării online.	Total de acord
	Activitățile online sunt mai eficiente decât cele clasice pentru dezvoltarea competențelor de socializare, de comunicare, de relaționare ale elevilor.	Total de acord
	Elevii sunt mai motivați să participe la activitățile online pentru că sunt atrași de calculatoare.	Abținere
	Dezvoltarea personalității elevilor este posibilă doar prin interacțiunea cu ceilalți elevi în sala de clasă.	Dezacord
	Evaluarea online favorizează obținerea unor note mari de către toți elevii, chiar și cei care nu depun eforturile necesare pentru a învăța.	Abținere
Elevi - C4	Activitățile online sunt mult mai atractive pentru elevi.	De acord
	Predarea online mi se pare cel puțin la fel de eficientă ca cea față în față.	De acord
	Utilizarea unor instrumente online de evaluare facilitează învățarea pentru că permite elevilor să știe instantaneu ce rezultate au avut la evaluări.	De acord
	Sunt teme / subiecte / materii care nu pot fi predate online într-un mod eficient.	De acord
	Activitățile online sunt mult mai atractive pentru elevi.	Abținere
	Predarea online este foarte eficientă pentru că elevi sunt familiarizați cu utilizarea calculatoarelor și dispozitivelor portabile.	De acord
Părinți - C4	Activitățile online sunt plictisitoare și monotone pentru elevi în lipsa interacțiunii directe cu colegii și profesorii.	Dezacord
	Evaluarea online favorizează obținerea unor note mari de către toți elevii, chiar și cei care nu depun eforturile necesare pentru a învăța.	Abținere
	Predarea online este foarte eficientă pentru că elevii sunt familiarizați cu utilizarea calculatoarelor și dispozitivelor portabile.	Dezacord

În urma analizei datelor colectate în 2021 (C4), au fost obținute perspective variate de la profesori, elevi și părinți privind eficiența și atractivitatea activităților online în învățământul preuniversitar din România. Aceste rezultate relevă câteva tendințe și aspecte cheie care pot contribui la îmbunătățirea calității educației.

Secțiunea *Importanța Variabilelor* /predictorilor/factorilor din modelul *Deep Learning* listează factorii care au pondere importantă în crearea modelelor. În Tabel 6.26. este prezentată o sinteză a acestor predictori, constatări și recomandări (corelate cu predictorii/factorii):

Tabel 6.26 Sinteză predictori, constatări și recomandări (set C4 – profesori, elevi, părinți)

Predictor	Constatări	Recomandări
Atractivitatea și interactivitatea activităților online	**Profesori:** Unii profesori sunt de acord că activitățile online sunt atractive pentru elevi, dar există și preocupări majore privind plictiseala și monotonia în lipsa interacțiunii directe. **Elevi:** Elevii par să găsească activitățile online mai atractive și consideră predarea online cel puțin la fel de eficientă ca cea față în față. **Părinți:** Părinții sunt în dezacord cu ideea că activitățile online sunt atractive, subliniind importanța interacțiunii directe.	**Dezvoltarea de activități online interactive** care să includă sesiuni de grup, discuții live și colaborări între elevi pentru a crește interactivitatea și atractivitatea. **Integrarea unor instrumente digitale** (platforme educaționale interactive și resurse multimedia) pentru a menține interesul și implicarea elevilor.
Eficiența predării online	**Profesori:** Există un acord că s-au făcut progrese mari în calitatea și rezultatele învățării online, dar există dezacorduri majore privind eficiența activităților online pentru dezvoltarea competențelor de socializare și comunicare. **Elevi:** Elevii consideră predarea online eficientă și apreciază utilizarea instrumentelor de evaluare online pentru feedback instantaneu. **Părinți:** Părinții sunt sceptici cu privire la eficiența predării online, considerând că interacțiunea directă este esențială pentru dezvoltarea personalității elevilor.	**Formare continuă pentru profesori** prin organizarea de sesiuni de formare continuă pentru profesori în utilizarea tehnologiilor educaționale și metodologiilor de predare online. **Evaluare continuă, adaptarea metodologiilor de predare în funcție de feedback-ul primit de la elevi și părinți și monitorizarea performanțelor**.
Impactul asupra socializării și dezvoltării personale	**Profesori și părinți:** Există un consens că activitățile online nu sunt eficiente pentru dezvoltarea competențelor de socializare și comunicare, subliniind importanța interacțiunii față în față. **Elevi:** Elevii au păreri împărțite, însă recunosc că anumite materii și subiecte sunt dificil de predat online.	**Implementarea unui model de învățare hibrid** care să combine avantajele învățământului online cu cele ale interacțiunilor față în față pentru a susține dezvoltarea personală și socializarea. **Dezvoltarea de programe de mentorat** și sesiuni de grup pentru a promova competențele de relaționare și comunicare în mediul online.

Predictor	Constatări	Recomandări
Evaluarea online	**Profesori și părinți:** Există preocupări că evaluarea online poate favoriza obținerea unor note mari de către elevi care nu depun eforturile necesare. **Elevi:** Elevii apreciază evaluările online pentru feedback-ul instantaneu și facilitarea învățării.	**Dezvoltarea și implementarea unor instrumente de evaluare online mai robuste și variate** care să reducă posibilitatea de fraudare și să evalueze mai eficient competențele elevilor. **Promovarea transparenței și a corectitudinii în evaluările online** prin utilizarea de algoritmi de detecție a plagiatului și a examinărilor asistate digital.

Aceste recomandări, bazate pe rezultatele cercetării, pot contribui semnificativ la îmbunătățirea calității învățământului preuniversitar din România prin abordarea provocărilor și valorificarea oportunităților oferite de tehnologiile educaționale moderne.

6.3.2 Educația online după 2 ani de la debutul pandemiei

În anul 2022 am colectat un set de date prin intermediul unui chestionar, cu scopul de a analiza evoluția percepțiilor diferitelor părți implicate în procesul educațional, la doi ani după începutul pandemiei. Setul de date rezultat, C5, a fost descris amănunțit în Capitolul 5 din teză.

Pentru a atinge obiectivele cercetării, este necesară identificarea celor mai adecvate modele de analiză a sentimentelor exprimate de cele trei categorii de respondenți cu privire la desfășurarea orelor online și impactul tehnologiilor digitale. Ulterior, din aceste modele, se va realiza identificarea factorilor determinanți care contribuie la manifestarea sentimentelor negative, cu scopul de a oferi recomandări pentru îmbunătățirea practicilor educaționale.

Pentru preprocesare, am înlocuit valorile colectate din chestionare cu șiruri de caractere simple, dar sugestive, astfel: *Total dezacord* - TD, *Dezacord* - D, *Abținere* - AB, *De acord* - A, *Total de acord* -TA.

Etapele parcurse sunt prezentate în Fig. 6.1 și Fig. 6.2, iar modelarea a fost realizată după procese similare celui descris în Fig. 6.3.

Tehnicile de clasificare considerate sunt similare celor abordate pentru seturile de date C1-C4: *Decision Tree, Random Forest, KNN, Naïve Bayes, SVM și Deep Learning*.

Rezultatele din Tabel 6.27 prezintă performanțele clasificatorilor considerați pentru răspunsurile profesorilor (C5), în două seturi de date: setul de date structurate și setul complet de date.

Tabel 6.27 Performanțele clasificatorilor pentru răspunsurile profesorilor (C5)

	Set date structurate				Set complet de date			
	Acurat [%]	**Prec [%]**	**Recall [%]**	**scor F [%]**	**Acurat [%]**	**Prec [%]**	**Recall [%]**	**scor F [%]**
DT	81,96	80,18	75,38	77,23	59,35	55,52	27,12	35,23

RF	**87,32**	89,12	78,59	83,39	59,68	51,84	33,57	40,25
NB	82,44	77,43	81,78	79,26	62,6	54,58	57,29	55,78
KNN	78,23	77,77	67,2	71,21	61,3	53,33	48,62	50,76
SVM	67,48	63,35	51,77	56,32	62,44	54,68	51,79	53,01
DL	**84,09**	81,39	80,27	80,4	53,97	45,91	70	55,24

Pentru setul de date structurate, *Random Forest* are cea mai mare acuratețe (87,32%), fiind urmat de *Deep Learning* cu 84,09%. *RF* se remarcă și prin cea mai mare precizie (89,12%) și scor F comparabil cu DL, ceea ce sugerează că acest model este foarte eficient în a face distincția între clase, fiind totodată aproape la fel de robust ca *DL*.

Pe setul complet de date modelele prezintă performanțe mult mai slabe, indicând o complexitate crescută a datelor care afectează precizia clasificării. *Naïve Bayes* se remarcă cu cea mai bună acuratețe (62,6%), sugerând că acest model este cel mai eficient în a identifica corect clasele pozitive, deși precizia este destul de scăzută (54,85%).

În Tabel 6.28 sunt prezentate rezultatele performanțelor clasificatorilor obținute pe seturile de date colectate de la elevi (C5-2022).

Tabel 6.28 Performanțele clasificatorilor pentru răspunsurile elevilor (C5)

	Set date structurate				**Set complet de date**			
	Acurat [%]	**Prec [%]**	**Recall [%]**	**scor F [%]**	**Acurat [%]**	**Prec [%]**	**Recall [%]**	**scor F [%]**
DT	85,89	89,49	84,46	86,83	84,32	82,22	91,86	86,65
RF	**89,83**	90,76	91,07	90,85	**88,98**	89,3	91,33	90,22
NB	87,85	92,25	85,47	88,63	74,31	73,91	83,96	78,39
KNN	82,76	86,79	81,88	84,15	81,91	86,37	81,11	83,37
SVM	69,62	68,24	84,68	75,49	67,5	66,95	81,85	73,57
DL	**87,86**	88,91	89,53	89,08	84,32	86,96	84,96	85,73

Pe setul de date structurate cele mai bune rezultate sunt obținute cu modelul *Random Forest* (89,83%), urmat îndeaproape de *Deep Learning* (87,86%).

Pe setul complet de date, acuratețea modelelor este în general mai mică decât pe setul structurat. *Random Forest* obține cea mai bună acuratețe, de 88,98 %. *Decision Tree* și *Deep Learning* mențin performanțe competitive, cu o acuratețe la egalitate, de 84,32%.

Deși toate modelele arată o creștere a recall-ului, *Decision Tree* și *Random Forest* obțin valori foarte ridicate, în jurul valorii de 91-92%. Aceasta sugerează o capacitate sporită de a recunoaște clasele pozitive.

Random Forest obține și cel mai bun scor F1, de 90,22%

Aceste rezultate indică faptul că, pentru analiza sentimentelor elevilor în contextul educațional, modelele *Random Forest* și *Deep Learning* oferă soluțiile cele mai eficiente, fiind capabile să gestioneze complexitatea datelor și să obțină un echilibru între precizie și recall.

În Tabel 6.29 sunt prezentate rezultatele performanțelor clasificatorilor obținute pe seturile de date colectate de la părinți (2022).

Tabel 6.29 Performanțele clasificatorilor pentru răspunsurile părinților (C5)

	Set date structurate				Set complet de date			
	Acurat [%]	**Prec [%]**	**Recall [%]**	**scor F [%]**	**Acurat [%]**	**Prec [%]**	**Recall [%]**	**scor F [%]**
DT	88,52	92,49	90,73	91,58	67,18	68,87	95,25	79,92
RF	**91,54**	90,63	97,97	94,12	67,96	68,85	97,4	80,66
NB	86,19	94,32	85,09	89,42	60,2	74,55	63,83	68,58
KNN	83,4	86,74	89,84	88,16	63,84	70,73	80,8	75,41
SVM	78,28	80,15	90,97	85,17	68,04	69,15	96,49	80,54
DL	**89,91**	93,85	91,29	92,54	68,11	69	97,29	80,73

Random Forest obține cele mai bune performanțe generale pe setul de date structurate, cu o acuratețe de 91,54%, un recall de 97,97% și un scor F de 94,12%. Acest aspect sugerează că *Random Forest* este cel mai eficient clasificator în identificarea corectă a clasei pozitive, menținând un echilibru optim între precizie și recall.

Deep Learning indică o performanță foarte bună, apropiată de cea a RF, cu o acuratețe de 89,91% și un scor F de 92,54%,

Pe setul complet de date performanțele scad semnificativ pentru majoritatea algoritmilor, ceea ce indică faptul că acest set de date este mai complex și poate conține mai mult zgomot.

Random Forest și *Deep Learning* continuă să performeze bine relativ la ceilalți algoritmi, cu o acuratețe de 67,18%, respectiv 68,11% și scor F de 80,66%, respectiv de 80,73%. Acești algoritmi își mențin capacitatea de a gestiona complexitatea crescută a datelor, deși performanța lor este afectată comparativ cu setul structurat.

Cu scopul de a ghida selectarea și optimizarea modelelor în funcție de tipul și structura datelor disponibile, concluzionez că *Random Forest* și *Deep Learning* sunt algoritmii cei mai robuști și eficienți în ambele seturi de date, gestionând bine atât datele structurate, cât și pe cele complete.

Scăderea generală a performanței pe setul complet de date subliniază importanța preprocesării și a gestionării complexității datelor pentru a menține acuratețea și fiabilitatea modelelor de clasificare.

Deși *Random Forest* a arătat performanțe ușor superioare față de *Deep Learning*, acesta din urmă este ales pentru interpretare datorită capacității sale de a gestiona și învăța din date complexe, potențialului său inovator și relevanței sale pentru cercetarea avansată în educație. Această alegere este susținută de următoarele considerente:

- *Deep Learning* este cunoscut pentru capacitatea sa de a modela relații complexe și nelineare în date, ceea ce este esențial în domeniul educațional, unde factorii care influențează rezultatele elevilor sau percepțiile acestora sunt adesea interdependenți și nu se pot modela ușor cu tehnici mai simple [56];
- Modelul Deep Learning are capacitatea de a învăța reprezentări de nivel înalt ale datelor prin straturi multiple de rețele neuronale, ceea ce poate fi extrem de util atunci când se analizează seturi de date complexe sau când se încearcă să se extragă cunoștințe din date [57];
- Deep Learning este mai robust la variațiile din date și poate oferi o performanță mai stabilă atunci când seturile de date sunt mari și complexe, cum este cazul în educație, unde datele pot include răspunsuri textuale, interacțiuni elev-profesor, și multe alte tipuri de variabile. Capacitatea modelului Deep Learning de a învăța automat caracteristici din datele brute (fără a necesita preprocesări exhaustive sau selecții manuale de caracteristici)

poate reduce riscul de a omite informații relevante care ar putea fi decisive pentru înțelegerea profundă a fenomenelor educaționale [58];

- Deep Learning este un subdomeniu al învățării automate care folosește rețele neuronale profunde pentru a modela relații complexe și pentru a învăța din datele disponibile. În analiza sentimentelor, Deep Learning a adus îmbunătățiri semnificative față de metodele tradiționale datorită capacității sale de a înțelege contexte complexe. Prin antrenarea pe seturi mari de date, modelele de Deep Learning pot învăța să recunoască tipare subtile în exprimarea sentimentelor și pot generaliza mai bine pe date noi și nestructurate [59];
- În loc să depindă de caracteristici construite manual (cum ar fi lista de cuvinte pozitive/negative), Deep Learning poate învăța automat reprezentări utile din datele brute, cum ar fi relațiile semantice dintre cuvinte și fraze [60].

Utilizarea Deep Learning pentru analiza sentimentelor profesorilor, elevilor și părinților din datele colectate prin chestionare reprezintă o abordare puternică și eficientă pentru a înțelege mai bine percepțiile și emoțiile implicate în procesul educațional. Această metodă permite o analiză detaliată și nuanțată a textului, oferind informații valoroase care pot fi utilizate pentru a îmbunătăți calitatea educației, pentru a personaliza experiențele de învățare și pentru a monitoriza constant starea de bine a tuturor participanților la procesul educațional [61].

Secțiunea *Importanța Variabilelor* /predictorilor/factorilor din modelul *Deep Learning* listează factorii care au pondere importantă în crearea modelelor.

În Tabel 6.30 sunt prezentați predictorii/factorii care au pondere semnificativă în inducerea modelelor de clasificare pentru setul de date C5.

Tabel 6.30 Predictorii/factorii care au pondere importantă în crearea modelelor pentru setul de date C5

Set de date	Predictori / factori	Opțiune respondent
Profesori – C5	Eficiența predării online	Total dezacord
	Progrese în școala online	Total dezacord
	Motivație elevi pentru școala online	Total dezacord
	Elevii fac progrese în școala online	Total de acord
	Directorii sunt buni manageri ai colegilor lor în mediul online	Total de acord
	Comunicarea online este dificilă pentru profesori	Total de acord
	Fondurile alocate sunt suficiente pentru educația online	Abținere
	Școala necesită digitalizare	Dezacord
	Elevii sunt interesați de învățare în mediul online	De acord
	Activitățile online sunt atractive pentru elevi	Total dezacord
Elevi – C5	Activitățile online sunt atractive pentru elevi	Total de acord
	Eficiența predării online	Total dezacord
	Motivație elevi pentru școala online	Total de acord
	Școala online este la fel de eficientă ca școala față în față	Total de acord
	Școala online poate completa eficient școala față în față	Total dezacord
	Elevii fac progrese în școala online	Total dezacord
Părinți – C5	Eficiența predării online	Total de acord
	Motivație elevi pentru școala online	Total dezacord
	Evaluarea online este subiectivă	Dezacord
	Activitățile online sunt atractive pentru elevi	De acord

	Progrese în școala online	Total dezacord
	Profesorii sunt bine pregătiți pentru predarea online	Abținere

În urma analizei datelor colectate în 2022 (C5), au fost obținute perspective variate de la profesori, elevi și părinți privind eficiența și atractivitatea activităților online în învățământul preuniversitar din România. Aceste rezultate relevă câteva tendințe și aspecte cheie care pot contribui la îmbunătățirea calității educației.

În Tabel 6.31 este prezentată o sinteză a predictorilor care au pondere importantă în crearea modelelor (*set de date C5 – profesori, elevi, părinți)*, constatări și recomandări (corelate cu predictorii):

Tabel 6.31 Sinteză predictori, constatări și recomandări (set de date C5 – *profesori, elevi, părinți*)

Predictor	Constatări	Recomandări
Eficiența predării online	**Profesori**: Majoritatea profesorilor nu consideră predarea online eficientă, ceea ce sugerează o percepție negativă puternică față de acest mod de educație. Această percepție negativă poate fi influențată de lipsa interactivității, dificultățile tehnice sau adaptabilitatea scăzută a elevilor la acest format. **Elevi**: Un număr semnificativ de elevi sunt în total dezacord cu ideea că predarea online este eficientă, indicând percepții negative asupra eficacității acestui mod de predare. **Părinți**: Părinții care sunt total de acord cu eficiența predării online sugerează o percepție pozitivă și încredere în acest format de educație. Aceasta indică faptul că, din perspectiva lor, predarea online reușește să îndeplinească obiectivele educaționale într-o manieră satisfăcătoare.	• formare continuă a profesorilor pentru îmbunătățirea competențelor digitale • Reevaluarea metodelor de predare online pentru a identifica aspectele care nu funcționează eficient și implementarea de strategii noi care să îmbunătățească interacțiunea și învățarea. • Oferirea de formare continuă pentru profesori pentru a le îmbunătăți competențele în utilizarea tehnologiilor și a platformelor de învățare online, astfel încât să crească eficiența predării. • Pentru a menține și îmbunătăți această percepție pozitivă, este important să se continue dezvoltarea și implementarea de metode inovatoare de predare online care să rămână interactive și eficiente. • Se recomandă extinderea accesului la resurse educaționale digitale și asigurarea formării continue a cadrelor didactice pentru a maximiza potențialul educației online.
Progrese în școala online	**Profesorii** consideră că elevii nu fac progrese semnificative în cadrul școlii online, ceea ce indică o problemă în modul de	• Este important să se dezvolte metode de evaluare mai adaptate la mediul online, care să ofere o măsurare mai precisă a progresului elevilor;

Predictor	Constatări	Recomandări
	evaluare sau în eficacitatea procesului de învățare online. **Părinții** care sunt în total dezacord cu faptul că elevii fac progrese în școala online exprimă îngrijorări serioase cu privire la eficacitatea acestui format de învățare. Aceasta sugerează că, din punctul lor de vedere, educația online nu reușește să asigure dezvoltarea academică adecvată a elevilor.	• Îmbunătățirea suportului educațional și personalizarea experienței de învățare în mediul online, prin utilizarea platformelor educaționale care monitorizează performanțele elevilor în timp real. • Este necesară o reevaluare a metodologiilor de predare și a conținutului livrat online pentru a identifica lacunele care împiedică progresul elevilor. • Implementarea de măsuri suplimentare, cum ar fi sprijinul individualizat, monitorizarea mai strictă a progresului și feedback-ul continuu, pentru a asigura că elevii sunt în mod constant stimulați și sprijiniți să avanseze în cadrul școlii online.
Motivație elevi pentru școala online	**Profesorii** sunt de părere că elevii nu sunt motivați în cadrul școlii online, ceea ce ar putea duce la o scădere a performanțelor academice și a participării active. **Elevii** care consideră că sunt motivați pentru școala online indică o atitudine pozitivă și o adaptare bună la acest format de educație. Un număr semnificativ de **părinți** consideră că elevii nu sunt motivați în cadrul școlii online. Aceasta sugerează o lipsă de implicare sau entuziasm din partea elevilor, ceea ce poate afecta negativ rezultatele educaționale.	• Implementarea unor strategii de motivare, cum ar fi activități de învățare pe bază de joc, recompense pentru participare și implicare, și crearea de comunități virtuale pentru a încuraja interacțiunea socială; • Implicarea părinților și crearea unui mediu de învățare acasă care să sprijine motivația elevilor. • Susținerea și consolidarea factorilor care contribuie la această motivație, cum ar fi utilizarea de metode de învățare atractive, personalizate și interactive. • Este esențial să se implementeze strategii pentru creșterea motivației elevilor, cum ar fi includerea activităților de învățare prin joc, recompense pentru participare și crearea unor programe care să fie mai atrăgătoare pentru elevi. • Colaborarea cu părinții pentru a identifica și aborda cauzele demotivării și pentru a susține elevii în adaptarea la formatul online de învățare.

Predictor	Constatări	Recomandări
Elevii fac progrese în școala online	Există și o parte a profesorilor care consideră că elevii fac progrese în cadrul școlii online, ceea ce sugerează că, în anumite condiții, acest mod de învățare poate fi eficient.	• Continuarea și dezvoltarea practicilor care au demonstrat succes, asigurând totodată adaptabilitatea la nevoile individuale ale elevilor; • Cercetarea mai detaliată a factorilor care contribuie la succesul acestor elevi pentru a le aplica în alte contexte educaționale.
Directorii sunt buni manageri ai colegilor lor în mediul online	Profesorii consideră că directorii gestionează bine echipa în mediul online, ceea ce sugerează un leadership eficient și o capacitate de adaptare la noile cerințe tehnologice.	• Continuarea susținerii directorilor în rolul lor de lideri, oferindu-le resurse și formare suplimentară pentru a-și menține și îmbunătăți abilitățile de management în mediul digital. • Împărtășirea bunelor practici și strategiilor de succes utilizate de acești directori pentru a le aplica la nivelul întregii instituții.
Comunicarea online este dificilă pentru profesori	Comunicarea online este percepută ca fiind dificilă de către profesori, ceea ce poate afecta negativ colaborarea și interacțiunea cu elevii și colegii.	• Formare continuă a profesorilor pentru a îmbunătăți abilitățile de comunicare digitală ale profesorilor; • Dezvoltarea unor protocoale clare de comunicare și a unor instrumente care să simplifice și să eficientizeze colaborarea online.
Fondurile alocate sunt suficiente pentru educația online	Profesorii sunt indecișI în privința suficienței fondurilor alocate pentru educația online, ceea ce sugerează incertitudine sau o lipsă de transparență în distribuția resurselor.	• Creșterea transparenței în alocarea și utilizarea fondurilor pentru educația online. • Evaluarea nevoilor specifice ale școlilor și profesorilor și redistribuirea resurselor în mod echitabil pentru a acoperi aceste nevoi.
Școala necesită digitalizare	Unii profesori nu sunt de acord cu necesitatea digitalizării școlii, ceea ce poate indica o rezistență la schimbare sau o percepție că metodele tradiționale sunt mai eficiente.	• Oferirea de informații și formare suplimentară despre beneficiile digitalizării pentru a schimba percepțiile negative și pentru a demonstra impactul pozitiv al tehnologiei în educație; • Implementarea graduală a digitalizării pentru a permite o adaptare mai ușoară și mai eficientă a profesorilor și elevilor.
Elevii sunt interesați de învățare în mediul online	Există o percepție pozitivă în rândul profesorilor că elevii sunt interesați de învățarea online, ceea ce sugerează că	• Consolidarea interesului elevilor prin diversificarea metodelor de predare online și includerea de activități interactive și atractive.

Predictor	Constatări	Recomandări
	acest mod de educație poate fi atractiv pentru unii elevi.	• Crearea de programe care să capitalizeze pe acest interes, dezvoltând module de învățare care să fie în același timp educative și captivante.
Activitățile online sunt atractive pentru elevi	**Profesorii** consideră că activitățile online nu sunt atractive pentru elevi, ceea ce indică o problemă în modul în care sunt concepute și implementate aceste activități. Un segment al **elevilor** consideră că activitățile online sunt atractive, ceea ce sugerează că aceste activități reușesc să capteze interesul elevilor și să le mențină atenția **Părinții** care sunt de acord că activitățile online sunt atractive pentru elevi indică faptul că, în general, aceste activități reușesc să capteze interesul și atenția elevilor. Aceasta sugerează că există componente ale educației online care sunt bine primite de către elevi.	• Conceperea de activități online cât mai interactive, relevante și antrenante pentru elevi; • Implicarea elevilor în procesul de design al activităților pentru a se asigura că acestea răspund intereselor și nevoilor lor. • Extinderea și diversificarea acestor activități online atractive pentru a include diferite metode interactive și multimedia, astfel încât să continue să fie captivante pentru un număr mai mare de elevi. • Încurajarea feedback-ului constant de la elevi pentru a ajusta și îmbunătăți activitățile online, menținând astfel nivelul lor de atractivitate. • Pentru a valorifica acest feedback pozitiv, se recomandă diversificarea activităților online și includerea unor metode interactive, precum simulările, jocurile educative și proiectele colaborative. • Monitorizarea continuă a feedback-ului de la elevi și părinți pentru a adapta și îmbunătăți constant activitățile online, asigurându-se că acestea rămân captivante și relevante.
Evaluarea online este subiectivă	**Părinții** care sunt în dezacord cu ideea că evaluarea online este subiectivă sugerează o percepție pozitivă față de obiectivitatea metodelor de evaluare utilizate în mediul online. Aceasta arată că părinții au încredere în corectitudinea și imparțialitatea acestor evaluări.	• Este important să se mențină și să se îmbunătățească în continuare practicile de evaluare online pentru a asigura obiectivitatea și transparența, inclusiv prin utilizarea de instrumente standardizate de evaluare. • Se recomandă organizarea de sesiuni de informare pentru părinți și elevi despre metodele de evaluare utilizate, pentru a menține încrederea în sistemul de evaluare online.
Profesorii sunt bine pregătiți	**Părinții** care s-au abținut de la a exprima o opinie clară privind pregătirea profesorilor pentru	• Este important să se îmbunătățească comunicarea și transparența privind formarea și pregătirea continuă a

Predictor	Constatări	Recomandări
pentru predarea online	predarea online arată o incertitudine sau o lipsă de cunoștințe în acest domeniu. Aceasta poate indica o comunicare insuficientă între școală și părinți cu privire la competențele digitale ale profesorilor.	profesorilor pentru predarea online. Școlile ar trebui să informeze regulat părinții despre eforturile depuse pentru a asigura pregătirea adecvată a cadrelor didactice. • Organizarea de sesiuni demonstrative sau prezentări care să arate părinților cum își desfășoară profesorii activitatea online și cum sunt utilizate noile tehnologii în predare, pentru a crește încrederea acestora în competențele profesorilor.

6.3.3 Cercetări privind evoluția factorilor determinanți ai opiniilor negative

Pandemia COVID-19 a determinat o tranziție rapidă și necesară către învățământul online, care a avut un impact semnificativ asupra percepțiilor și opiniilor participanților la procesul educațional. Pe măsură ce perioada pandemiei s-a extins, s-au schimbat și percepțiile asupra eficienței și eficacității acestui tip de educație. Acest subcapitol explorează factorii care au determinat opinii negative în rândul profesorilor, elevilor și părinților, pe baza chestionarelor aplicate în anii 2020, 2021 și 2022.

În primele luni de la debutul pandemiei, în 2020, factorii determinanți ai percepțiilor negative au fost legați în principal de ***lipsa unui cadru legal solid pentru educația online, mediul în care se desfășura procesul educațional și diferențele în competențele IT ale profesorilor***. Profesorii au simțit nevoia unor reglementări clare și eficiente pentru susținerea educației online. Pe de altă parte, diferențele semnificative în nivelul competențelor digitale ale profesorilor au condus la dificultăți majore în adaptarea la noile tehnologii. Astfel, schimbările drastice în modalitatea de predare au fost întâmpinate cu scepticism, profesorii manifestând preferința pentru învățământul tradițional, datorită provocărilor tehnologice și a infrastructurii insuficiente.

Acești factori au contribuit la o percepție negativă generalizată asupra educației online în stadiul său incipient.

În 2021, după un an de învățământ online, percepțiile negative s-au concentrat pe atractivitatea activităților online și eficiența acestora. Deși o parte dintre profesori și elevi au recunoscut progrese în creșterea calității educației online, alții au considerat că ***activitățile online sunt monotone și că dezvoltarea competențelor de socializare este mai eficientă în învățământul față în față***. Părinții, pe de altă parte, au manifestat îngrijorări legate de ***subiectivitatea evaluărilor și de lipsa interacțiunii directe***.

Percepția elevilor a fost divizată, unii considerând activitățile online atractive și predarea eficientă, în timp ce alții au identificat limitări semnificative, în special în ceea ce privește predarea anumitor subiecte și interacțiunea socială.

În acest stadiu, educația online a început să fie văzută cu mai multă ambivalență, recunoscându-se atât avantajele, cât și limitările acesteia.

După doi ani de educație online, în 2022, s-au observat opinii puternic polarizate. Factorii care au determinat opiniile negative au devenit mai clar conturați, reflectând o experiență acumulată și un context mai stabilizat. Profesorii și părinții au continuat să fie împărțiți, unii exprimând un total ***dezacord față de eficiența predării online***, în timp ce alții și-au menținut

acordul. Aceasta sugerează o **polarizare a opiniilor** bazată pe experiențe individuale. În ceea ce privește progresul în școala online, profesorii și părinții au exprimat un ***total dezacord față de ideea că elevii au făcut progrese semnificative în școala online***, ceea ce indică o continuare a provocărilor legate de adaptarea la acest tip de educație.

Lipsa motivației elevilor a fost recunoscută ca un obstacol major în succesul educației online, subliniind necesitatea de a găsi metode mai eficiente pentru a menține implicarea elevilor. În ciuda unor îmbunătățiri percepute, mulți profesori au continuat să considere ***activitățile online ca fiind neatractive pentru elevi***, ceea ce a contribuit la menținerea unui sentiment general de insatisfacție.

Analiza evoluției factorilor determinanți ai opiniilor negative de-a lungul a trei ani de educație online (2020-2022) relevă o adaptare treptată, dar și o persistență a provocărilor majore. Deși au fost făcute progrese, în special în dezvoltarea competențelor digitale și în adaptarea materialelor didactice, eficiența și atractivitatea educației online rămân puncte nevralgice.

În timp ce unele percepții negative au rămas constante, cum ar fi îngrijorările legate de eficiența predării și de motivația elevilor, s-au făcut progrese semnificative în alte domenii, cum ar fi recunoașterea importanței competențelor manageriale ale directorilor și a avantajelor oferite de digitalizare. Polarizarea opiniilor, atât între diferitele categorii de respondenți, cât și în cadrul aceleiași categorii, indică nevoia unor intervenții personalizate, care să răspundă mai bine nevoilor și așteptărilor fiecărui grup implicat în procesul educațional.

Aceste constatări sugerează că, deși educația online a devenit o componentă esențială a sistemului educațional modern, rămân provocări semnificative care trebuie abordate pentru a transforma complet percepțiile negative în oportunități de îmbunătățire a experienței de învățare.

În concluzie, analiza evoluției factorilor determinanți ai opiniilor negative asupra educației online între 2020 și 2022 relevă necesitatea unor intervenții continue și adaptabile.

6.3.4 Recomandări

În urma cercetărilor efectuate, cu scopul de a atenua percepțiile negative și de a contribui la îmbunătățirea calității educației, pregătind astfel terenul pentru o integrare eficientă și echitabilă a educației online în sistemul educațional, am sintetizat cei mai importanți factori ai opiniilor negative (Tabel 6.32).

Tabel 6.322 Sinteza celor mai importanți factori determinanți ai opiniilor negative și recomandări

Factorii determinanți ai opiniilor negative	Recomandări
Lipsa unui cadru legal și a unei pregătiri neadecvate a profesorilor a fost un factor major în generarea opiniilor negative.	• crearea unui cadru legal care să reglementeze eficient educația online; • dezvoltarea de politici și programe de formare care să susțină competențele IT ale profesorilor; • implementarea unor programe de formare continuă pentru profesori, care să le permită să utilizeze eficient noile tehnologii și să îmbunătățească atractivitatea și eficiența lecțiilor online;

Persistența opiniilor negative în rândul profesorilor și părinților a fost influențată de percepțiile privind faptul că activitățile online sunt monotone și a lipsesc copiii de interacțiunea fizică.	• diversificarea activităților didactice online și implementarea unor metode interactive care să stimuleze participarea activă și socializarea elevilor; • • crearea unor strategii didactice inovative, care să combine avantajele învățământului online cu necesitatea interacțiunii sociale și a învățării active; • dezvoltarea unor metode de motivare care să capteze interesul elevilor și să încurajeze participarea activă în procesul de învățare online.
Persistența opiniilor negative în rândul profesorilor și părinților indică o rezistență față de educația online, în principal din cauza percepției că acest format nu oferă aceleași beneficii ca învățământul tradițional.	• este necesară o reevaluare a metodelor de predare și de evaluare online, punând un accent mai mare pe personalizarea învățării și pe asigurarea unui feedback constant pentru a monitoriza și sprijini progresul elevilor; • revizuirea instrumentelor de evaluare online pentru a asigura corectitudinea și obiectivitatea notării, reducând astfel percepția negativă legată de subiectivitatea evaluărilor online.

Analiza evoluției factorilor determinanți ai opiniilor negative privind educația online dezvăluie atât provocările, cât și oportunitățile pe care această formă de învățare le aduce. Deși percepțiile asupra educației online au fost inițial sceptice și adesea negative, există un potențial enorm pentru îmbunătățire prin adoptarea unor soluții inovatoare care să răspundă nevoilor specifice ale elevilor, profesorilor și părinților.

Deși s-au făcut progrese semnificative, există încă numeroase provocări care necesită soluții inovatoare. Pentru a transforma aceste provocări în oportunități și pentru a crea un mediu educațional online care să fie cu adevărat eficient, atractiv și accesibil pentru toți, este nevoie și de platforme educaționale vizionare.

O platformă educațională revoluționară, concepută pentru a răspunde în mod holistic nevoilor elevilor, profesorilor și părinților ar combina tehnologiile de ultimă generație cu o abordare personalizată a educației, urmărind nu doar transmiterea de cunoștințe, ci și cultivarea motivației, a competențelor sociale și a abilităților critice. Cu o astfel de platformă, recunoscută de Ministerul Educației, educația online nu doar că ar depăși limitările identificate în aceste cercetări, dar ar deveni un pilon central al educației moderne, capabil să formeze generații de elevi bine pregătiți pentru viitor.

Pe măsură ce educația online continuă să evolueze, este esențială abordarea nu doar a provocărilor tehnice și pedagogice, ci și reinventarea modului în care această formă de educație este livrată.

7 Contribuțiile tezei și diseminarea rezultatelor

Cercetările desfășurate pe durata studiilor doctorale au avut ca finalitate o serie de contribuții.

Lucrarea contribuie la avansul cercetărilor în domeniu printr-o suită de contribuții, atât teoretice cât și practice.

T1: ***O analiză comparativă*** a modelelor asociate procesului de descoperire a cunoștințelor din date, și relevarea caracteristicilor distinctive. (**Capitolul 2**).

T2: ***Analiza tendințelor*** privind utilizarea practică a modelelor KDD, începând cu anul 2007 până în prezent. (**Capitolul2**)

T3: ***Conceptualizarea unor probleme*** din domeniul educațional prin formularea a 23 de întrebări al căror răspuns poate fi furnizat prin explorarea datelor educaționale **(Capitolul 2)**

T4: O investigare a stadiului actual privind cercetările și utilizarea aplicațiilor de EDM în Uniunea Europeană pentru care am realizat o analiză sistematică a literaturii de specialitate (**Capitolul 4**). Procesul de analiză a fost realizat pe baza metodologiei Kitchenham, în principalele baze de date academice: Scopus, Science Direct și IEEE Xplore. Pentru a asigura relevanța și actualitatea studiului, publicațiile studiate vizează perioada 2013-2023. Au fost identificate 847 de studii din care, după aplicarea criteriilor de includere și de excludere au fost selectate 17 studii. Sinteza rezultatelor analizei sistematice a permis obținerea unei imagini clare a stadiului actual al cercetărilor în domeniul EDM la nivelul Uniunii Europene, identificând tendințele majore, lacunele, provocările și oportunitățile pentru viitor.

T5: ***Propunerea unor direcții de cercetare în domeniul EDM în România, ca parte a Uniunii Europene.*** Concluziile analizei sistematice a literaturii de specialitate privind stadiul actual al Educational Data Mining la nivelul Uniunii Europene, au relevat câteva lipsuri privind cercetările pe alte direcții decât predicția performanțelor funcție de comportamentul de învățare. Pornind de la aceste constatări am propus câteva direcții viitoare de cercetare având ca țintă învățământul preuniversitar din România:

- investigarea influenței factorilor non-cognitivi, precum motivația, atitudinea și implicarea, asupra performanței academice;
- investigarea diferențelor de performanță academică între diverse regiuni și școli din România;
- dezvoltarea de modele predictive pentru a anticipa traiectoriile de carieră ale elevilor pe baza performanțelor academice și a intereselor lor;
- utilizarea data mining pentru a identifica elevii cu abilități și talente deosebite și pentru a dezvolta programe de suport pentru aceștia;
- investigarea impactului activităților extracurriculare asupra dezvoltării cognitive și sociale a elevilor.

T6: ***Elaborarea de recomandări*** privind acțiunile necesare de întreprins așa cum rezultă din identificarea factorilor care influențează în mod negativ procesul de învățare (**Capitolul 6)**.

T7: ***Măsurarea consistenței interne a itemilor*** din chestionarele utilizate în teză prin calcularea coeficientului alpha-Cronbach, folosind software-ul SPSS. Valori obținute sunt cuprinse între 0,736 și 1, ceea ce atestă fiabilitatea și acuratețea instrumentelor utilizate în cercetare (**Capitolul 5)**.

P1: ***Proiectarea a 6 chestionare*** destinate grupului țintă profesori, elevi și părinți și implementarea acestora în Google Forms (**Capitolul 5).**

P2: ***Coordonarea procesului de distribuție a chestionarelor către respondenți*** (**Capitolul 5).** În perioada aprilie 2020 – februarie 2024 au fost distribuite online, la nivel național, 6 chestionare cu scopul de a investiga aspecte privind educația din învățământul preuniversitar din România.

P3: ***Colectarea datelor din chestionare, salvarea acestora în fișiere Excel și analiza datelor colectate*** din punct de vedere al tipului și calității, precum și al reprezentativității eșantioanelor de respondenți. (**Capitolul 5**) Deoarece sondajele au presupus participarea pe bază de voluntariat, se consideră reprezentative eșantioanele formate din răspunsurile a 956 de profesori (2020), 1088 de elevi (2020), 784 de părinți (2020), 7701 de răspunsuri de la cele 3 categorii în anul 2021, 2612 de răspunsuri în 2022 și 1515 răspunsuri culese în 2024.

P4: ***Proiectarea și implementarea proceselor de inducție a modelelor și evaluare a performanțelor, precum și analiza rezultatelor obținute*****.** În cadrul cercetărilor noastre în domeniul Educational Data Mining, am implementat procese complexe care implică inducția modelelor, evaluarea performanțelor acestora și analiza rezultatelor obținute. (**Capitolul 6**)

P5: ***Proiectarea și implementarea unor procese de text mining pentru analiza textelor în limba română.*** (**Capitolul 6**)

P6: ***Implementarea principiilor Open Science în cercetarea educațională*****.**
Uniunea Europeană promovează *Open Science* prin directive precum *Planul S* și programele *Orizont 2020* și *Orizont Europa*. Aceste inițiative vizează accesibilizarea și transparența cercetării științifice prin încurajarea în scopul publicării cu acces deschis, partajarea datelor de cercetare conform principiilor FAIR, dezvoltarea infrastructurii digitale și formarea cercetătorilor.

Pentru creșterea impactului și vizibilității cercetărilor, mi-am propus să postez public datele colectate, într-un repository de date științifice susținut de CERN și integrat în European Open Science Cloud (EOSC) [51] Acestea vor putea fi descărcate în scopul procesării de pe platforma Zenodo. [52] care este o platformă robustă și accesibilă, destinată stocării, partajării și publicării datelor de cercetare din orice domeniu științific. Platforma oferă spațiu de stocare gratuit și asigură vizibilitatea datelor. Platforma Zenodo suportă diverse formate de date și facilitează integrarea cu alte infrastructuri de date științifice, asigurând interoperabilitatea și accesul deschis. Această alegere reflectă angajamentul meu pentru promovarea accesului liber la cunoaștere și pentru contribuția la dezvoltarea unui ecosistem de cercetare colaborativ și transparent.

În cadrul tezei mele de doctorat am decis să implementez principiile *Open Science* pentru a crește impactul și vizibilitatea cercetărilor mele. Contribuția mea specifică include următoarele aspecte:

- publicarea datelor cu acces deschis – toate datele colectate în cadrul cercetării mele vor fi postate public pe platforma *Zenodo*, o infrastructură digitală dezvoltată pentru stocarea și partajarea datelor de cercetare; datele vor fi disponibile pentru descărcare și procesare, permițând altor cercetători să valideze rezultatele mele, să extindă cercetarea sau să utilizeze datele în studii conexe;
- conformarea cu principiile *FAIR* (Findable Accessible Interoperable Reusable) – datele vor fi organizate și etichetate astfel încât să fie ușor de găsit, accesat, utilizat în alte platforme, sau contexte de cercetare;

- prin publicarea pe platforma *Zenodo* îmi aduc contribuția la dezvoltarea și consolidarea infrastructurii digitale necesare pentru *Open Science* prin îmbunătățirea accesului la date prin intermediul unor platforme recunoscute la nivel european, aspect ce poate facilita colaborarea internațională și interdisciplinară;
- prin exemplul personal și prin diseminarea rezultatelor și metodologiei utilizate, urmăresc să inspir și să încurajez alți cercetători să adopte practicile *Open Science*;
- creșterea impactului cercetării prin publicarea datelor cu acces deschis, aspect ce va facilita utilizarea acestora de către un număr mai mare de cercetători, crescând astfel impactul și vizibilitatea cercetărilor mele.

7.1 Diseminarea rezultatelor

Rezultatele cercetărilor au fost diseminate prin publicarea a 12 lucrări științifice indexate WOS, IEEE și BDI. Acestea au fost publicate la nivel național și internațional în jurnale sau în volumele unor conferințe la care au fost prezentate.

Două lucrări (A11, A12) au fost prezentate doar în cadrul conferințelor, publicarea făcându-se în volume de rezumate.

A1: C. Simionescu, M. Danubianu, D. Marcu, C. O. Turcu, *Online learning after one year of digital schooling in Romania – a survey*, IJCSNS International Journal of Computer Science and Network Security, VOL.21 No.12, December 2021 https://doi.org/10.22937/IJCSNS.2021.21.12.99 - **indexat Web of Science**, **JCI 0.09, Quartile Q4, Accession Number WOS:000755171400005**

A2: Corina Simionescu, Mirela Danubianu, Bogdanel Constantin Gradinaru and Marius Silviu Maciuca, *Educational Data Mining in European Union – Achievements and Challenges: A Systematic Literature Review*, International Journal of Advanced Computer Science and Applications (IJACSA), 15(3), 2024. http://dx.doi.org/10.14569/IJACSA.2024.0150386 - **indexat WoS, Quartile Q4, JCI 0.18, Accession Number WOS:001300899000001**

A3: C. Simionescu, M. Danubianu, A.-L. Bărîlă and B. C. Grădinaru, *Factors Influencing Romanian Teachers' Choice of IT Training to Avoid Issues Raised by Online Education: A Data Mining Approach*, 2024 International Conference on Development and Application Systems (DAS), Suceava, Romania, 2024, pp. 199-204, doi: 10.1109/DAS61944.2024.10541222. **IEEE Digital Library**

A4: **Simionescu, C.,** Marcu, D., & Măciucă, M. S. . (2024). *Toward Better Education Quality through Students' Sentiment Analysis Using AutoML*. BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 15(2), 320-343. https://doi.org/10.18662/brain/15.2/578, (**în curs de indexare WoS**)

A5: Daniela Marcu, Mirela Danubianu, Adina Bărîlă, **Corina Simionescu**, Stefan cel Mare University of Suceava, Romania, *Algorithms for Classifying the Results at the Baccalaureate Exam - Comparative Analysis of Performances*, IJCSNS International Journal of Computer Science and Network Security, VOL.21 No.8, August 2021 http://paper.ijcsns.org/07_book/202108/20210805.pdf,
DOI: https://doi.org/10.22937/IJCSNS.2021.21.8.5 **– indexat Web of Science, JCI 0.09, Quartile Q4, Accession Number WOS:000697025200005**

A6: Maciuca, Marius, Danubianu, Mirela, **Simionescu, Corina.** (2022). *Tendencies in the use of Big Data analytics at a global level*. 155-160. 10.1109/DAS54948.2022.9786116. **IEEE Digital Library**

A7: C. Simionescu, M. Danubianu, D. Marcu (2020), *Analysis of online education romanian schools due to covid-19 pandemics and areas of improvement*, ICERI2020 Proceedings, pp. 3523-3529, 13th annual International Conference of Education, Research and Innovation, Online Conference. 9-10 November, 2020. ISBN: 978-84-09-24232-0 / ISSN: 2340-1095, https://library.iated.org/view/SIMIONESCU2020ANA, https://doi.org/10.21125/iceri.2020.0787

A8: **Simionescu, C.**, Danubianu, M., & Maciuca, M. S. (2023). How Data Mining and Artificial Intelligence can Contribute to Increasing Academic Performance. Didactica Danubiensis, 3(1), 72–85. https://dj.univ-danubius.ro/index.php/DD/article/view/2467

A9: Corina Simionescu, Mirela Danubianu, Corneliu-Octavian Turcu, *Data mining in educational data – useful tool for sustainable learning development*, EIRP Proceedings of the INTERNATIONAL CONFERENCE European Integration - Realities and Perspectives 16th Edition, 14-15 may 2021, p 349-353, ISSN: 2067–9211 https://dp.univ-danubius.ro/index.php/EIRP/article/view/212/194

A10: Marcu, D., Danubianu M., **Simionescu C.** (2021), *Comparative analysis of predictve models on online education in context of covid-19 – A case study*, INTED2021 Proceedings, pp. 4403-4412, 15th International Technology, Education and Development Conference, Online Conference. 8-9 March, 2021. ISBN: 978-84-09-27666-0 / ISSN: 2340-1079 https://library.iated.org/view/MARCU2021COM, https://doi.org/10.21125/inted.2021.0899

A11: **Simionescu C.,** Danubianu M., Turcu C., *Study on online education in Romania during the Covid-19 pandemic*, 29 june 2021 - 11th International Conference The Danube - Axis of European Identity, Universitatea DANUBIUS Galati

A12: **Simionescu C.**, Danubianu M., Turcu C., Data Mining, *The benefits of data extraction from knowledge in the educational field*, 29 june 2021 - 11th International Conference The Danube - Axis of European Identity, Universitatea DANUBIUS Galati

7.1.1 Participarea cu prezentări în cadrul conferințelor internaționale

C1: ICERI2020 (13th annual International Conference of Education, Research and Innovation, 9th - 10th of November, 2020, Spain - virtually)

C2: INTED2021 (15th annual International Technology, Education and Development Conference, 8th-9th of March, 2021, Spain, Valencia - virtually)

C3: INTERNATIONAL CONFERENCE European Integration - Realities and Perspectives 16th Edition, 14-15 may 2021, Galați, România - virtually

C4:11th International Conference The Danube - Axis of European Identity, 29 june 2021, Galați, România – virtually

C5: INTERNATIONAL CONFERENCE Education in the Digital Era 2 nd Edition, Galati, July 27–28, 2023

C6: 2024 International Conference on Development and Application Systems (DAS), Suceava, Romania, 23-25 may 2024

7.1.2 Implicarea în proiecte de cercetare

Proiect ANTREPRENORDOC, perioada 1 mai 2021 – 30 aprilie 2022: Excelența academică și valori antreprenoriale - sistem de burse pentru asigurarea oportunităților de formare și dezvoltare a competențelor antreprenoriale ale doctoranzilor și postdoctoranzilor, cofinanţat din Fondul Social European prin Programul Operaţional Capital Uman, 2014-2020, Contract nr. 36355/23.05.2019 POCU/380/6/13 - Cod SMIS: 123847

7.2 Concluzii și dezvoltări viitoare

Această teză de doctorat analizează, prin metode și tehnici de data mining, eficiența procesului educațional online comparativ cu cel față în față, din perspectiva profesorilor, elevilor, dar și al părinților. Seturile de date supuse cercetării sunt originale și corespund perioadei de școală online desfășurată în perioada pandemiei de Covid 19 și după.

În cadrul tezei mele de doctorat privind utilizarea tehnicilor și metodelor de data mining în educație, cercetarea s-a concentrat pe analiza a cinci seturi de date originale (C1-C5) - colectate de la profesori, elevi și părinți din România. Aceste seturi de date au oferit o bază solidă pentru identificarea și aplicarea diverselor tehnici de data mining, conducând la concluzii valoroase referitoare la impactul acestor metode asupra mediului educațional.

Al șaselea set de date, denumit C6, va fi inclus în direcțiile viitoare de cercetare. Acest set de date, C6, este menit să extindă și să aprofundeze analiza realizată până acum, oferind o perspectivă suplimentară asupra unor aspecte cheie, precum personalizarea învățării, analiza comportamentului elevilor și evaluarea progresului educațional în timp. Având în vedere complexitatea și volumul mare de date din acest set, analiza acestuia va fi tratată în studii ulterioare pentru a permite o evaluare mai detaliată și riguroasă.

Comparația diferitelor instrumente și tehnici prezentate în acest studiu sunt coroborative și concludente și sugerez că în contextul educațional actual sunt recomandate studii suplimentare pentru fiecare dintre metodele de data mining, luând în considerare mai multe standarde pentru a stabili tehnicile cu mai multă acuratețe.

Înțelegerea comportamentului profesorilor, elevilor și a modului în care învață (cu toții) poate ajuta managementul educațional să îmbunătățească programele de studiu actuale și practica educațională în general. Prin analiza datelor educaționale, precum și prin analiza importanței influenței factorilor/predictorilor, diverse modele de data mining ar putea fi utilizate ca suport pentru luarea deciziilor în educație, contribuind astfel la studii de succes și la îmbunătățirea calității educației.

Beneficiile și aplicațiile explorării de date educaționale sunt numeroase. Există lucrări de cercetare și studii privind utilizarea și aplicațiile tehnicilor de data mining în educație care abordează: îmbunătățirea procesului de studiu, îmbunătățirea finalizării cursurilor, sprijinirea studenților în selecția cursurilor, profilarea studenților, găsirea de probleme care duc la abandonul școlar, direcționarea studenților, dezvoltarea curriculumului, predicția performanței și ca sprijin pentru luarea deciziilor.

Concluzia este că școala trebuie să se schimbe pentru a se adapta la cerințele prezente, dar şi pentru a răspunde viitoarei piețe a muncii; această schimbare ar trebui să înceapă imediat şi ar trebui să înceapă cu schimbarea mentalității profesorilor. Această schimbare ar putea fi modelată prin formare continuă motivată de dorința lor de a deveni profesorii viitorului.

Susțin ideea că pentru o mai bună învățare este nevoie de eforturi comune din partea tuturor celor trei actori educaționali: profesori, elevi, părinți, iar dezvoltarea abilităților digitale, alături de înțelegerea emoțiilor celor implicați în educație sunt factori cheie pentru atingerea acestui obiectiv.

Deși s-au făcut progrese semnificative, există încă numeroase provocări care necesită soluții inovatoare. Pentru a transforma aceste provocări în oportunități și pentru a crea un mediu educațional online care să fie cu adevărat eficient, atractiv și accesibil pentru toți, este nevoie și de platforme educaționale vizionare.

O altă direcție de cercetare este constituirea unei platforme educaționale revoluționară, concepută pentru a răspunde în mod holistic nevoilor elevilor, profesorilor și părinților ar combina tehnologiile de ultimă generație cu o abordare personalizată a educației, urmărind nu doar transmiterea de cunoștințe, ci și cultivarea motivației, a competențelor sociale și a abilităților critice. Cu o astfel de platformă, recunoscută de Ministerul Educației, educația online nu doar că ar depăși limitările identificate în aceste cercetări, dar ar deveni un pilon central al educației moderne, capabil să formeze generații de elevi bine pregătiți pentru viitor.

Această platformă va combina tehnologiile de ultimă generație cu o abordare personalizată a educației, urmărind nu doar transmiterea de cunoștințe, ci și cultivarea motivației, a competențelor sociale și a abilităților critice. Astfel, pot fi utilizați algoritmi de inteligență artificială pentru a personaliza experiența de învățare a fiecărui elev, adaptând conținutul, ritmul și metodele de predare la nevoile individuale. Fiecare elev ar beneficia de un traseu educațional unic, care să îi permită să progreseze în ritmul propriu, maximizând astfel eficiența și implicarea. Platforma va include instrumente de evaluare inteligente care să ofere feedback instant, obiectiv și detaliat. Aceste evaluări ar putea folosi analize predictive pentru a anticipa domeniile în care elevii ar putea întâmpina dificultăți, oferind astfel intervenții personalizate în timp util. Cu scopul de a transforma învățarea într-o experiență captivantă și motivațională, platforma ar trebui să includă elemente de învățare prin joc. Elevii vor putea debloca realizări, participa la competiții educative și câștiga recompense virtuale care să îi motiveze să își atingă obiectivele de învățare. În plus, crearea de comunități virtuale unde elevii, profesorii și părinții pot interacționa, colabora și împărtăși resurse va încuraja interacțiunea socială, schimbul de idei și sprijinul reciproc, contribuind la construirea unui sentiment de apartenență și suport.

Pentru a face învățarea și mai captivantă, această platformă va putea integra tehnologii de realitate virtuală și augmentată, permițând elevilor și profesorilor să exploreze medii virtuale interactive, să participe la experimente simulate și să învețe prin experiențe directe.

Cum astăzi comunicarea și educația au devenit globale și se petrec în timp real, școala trebuie să supravegheze, în colaborare cu instituțiile specializate acest proces educațional.

Data mining nu este un panaceu universal capabil să rezolve orice problemă. În fapt, aportul său se rezumă la un număr limitat de acțiuni: clasificarea, estimarea, predicția, gruparea, analiza grupărilor, dar care, folosite în mod adecvat, se pot dovedi extrem de utile pentru numeroase probleme și situații din domeniul decizional care pot conduce spre îmbunătățirea resurselor din învățământul preuniversitar din România.

Cu ajutorul data mining se pot obține informații prețioase legate de identificarea unor oportunități, găsirea unor metode de rezolvare pentru probleme complexe etc.

Toate acestea fac din tehnicile de explorare a datelor un instrument valoros pentru educație și un domeniu în care se fac continuu cercetări.

Mulțumiri

Exprim sincere mulțumiri doamnei ***prof. univ. dr. ing. Mirela Danubianu*** pentru sprijinul și îndrumarea oferite în cercetarea mea doctorală. Profesionalismul, cunoștințele și dedicarea sa mi-au fost surse de inspirație și motivare.

Sunt recunoscătoare pentru oportunitatea de a avea un mentor de o asemenea valoare academică și umană!

*

„Această lucrare a fost realizată în cadrul proiectului Excelența academică și valori antreprenoriale - sistem de burse pentru asigurarea oportunităților de formare și dezvoltare a competențelor antreprenoriale ale doctoranzilor și postdoctoranzilor (ANTREPRENORDOC), cofinanţat din Fondul Social European prin Programul Operaţional Capital Uman, 2014-2020, Contract nr. 36355/23.05.2019 POCU/380/6/13 - Cod SMIS: 123847.”

„This work is supported by the project ANTREPRENORDOC, in the framework of Human Resources Development Operational Programme 2014-2020, financed from the European Social Fund under the contract number 36355/23.05.2019 HRD OP /380/6/13 – SMIS Code: 123847.”

Bibliografie

[1]. Danubianu, Mirela & Pentiuc, Stefan. (2013). Data Dimensionality Reduction Framework for Data Mining. Electronics and Electrical Engineering. 19. 10.5755/j01.eee.19.4.2043.

[2]. Bienkowski M., Feng. M. &Means, B., Enhancing Teaching and Learning through Educational Data Miningand Learning Analytics: An Issue Brief [M]. Washington, D.C, 2012

[3]. Boicea, A., Truică, C. O., Rădulescu, F., & Buşe, E. C. (2018). Sampling strategies for extracting information from large data sets. Data & Knowledge Engineering, 115, 1-15.

[4]. Fayyad, Usama & Piatetsky-Shapiro, Gregory & Smyth, Padhraic. (2000). Knowledge Discovery and Data Mining: Towards a Unifying Framework.

[5]. Gullo, Francesco. (2015). From Patterns in Data to Knowledge Discovery: What Data Mining Can Do. Physics Procedia. 62. 18–22. 10.1016/j.phpro.2015.02.005.

[6]. Shivali, Joni Birla, Gurpreet, 2015, Knowledge Discovery in Data-Mining, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NCETEMS – 2015 (Volum 3 – Issue 10)

[7]. S. Ohsuga, Difference between data mining and knowledge discovery - a view to discovery from knowledge-processing, 2005 IEEE International Conference on Granular Computing, Beijing, China, 2005, pp. 7-12 Vol. 1, doi: 10.1109/GRC.2005.1547224.

[8]. Sarker, Ruhul & Abbass, Hussein & Newton, C.. (2000). Introducing Data Mining and Knowledge Discovery. 10.4018/9781930708266.ch001.

[9]. Tan, Steinbach, Kumar – Introduction to Data Mining, 2004

[10]. Ana Azevedo, M.F. Santos, KDD, SEMMA and CRISP-DM: a parallel overview, In Proceedings of the IADIS European Conference on Data Mining 2008

[11]. Danubianu, M., Pentiuc, S. G., & Danubianu, D. M. (2014). Data dimensionality reduction for data mining: a combined filter-wrapper framework. International Journal Of Computers Communications & Control, 7(5), 824-831.

[12]. M. Danubianu, A data preprocessing framework for students' outcome prediction by data mining techniques. In 19th International Conference on System Theory, Control and Computing (ICSTCC), p. 836-841, 2015

[13]. Friedman, J. H. (1998). Data Mining and Statistics: What's the connection?. Computing science and statistics, 29(1), 3-9.

[14]. https://www.oracle.com/ro/artificial-intelligence/machine-learning/what-is-machine-learning/

[15]. J. Stefanowski, Data Mining- Evaluation of Classifiers, Institute of Computing Sciences Poznan University of Technology, Poland, 2010

[16]. Han, J., Pei, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann

[17]. Srivastava, Durgesh & Bhambhu, Lekha. (2010). Data classification using support vector machine. Journal of Theoretical and Applied Information Technology. 12. 1-7.

[18]. Laviniu Aurelian Bădulescu, Arborele de decizie – model predictiv în data mining, Electrotehnica, electronica, automatica, 53 (2005), nr. 4

[19]. Mohammed J. Zaki, Wagner Meira Jr., Data mining and analysis. Fundamental Concepts and Algorithms, Cambridge University Press, ISBN 978-0-521-76633-3, 2014
[20]. Er-raji, Naoufal & Benabbou, Faouzia & Danubianu, Mirela & Zaouch, Amal. (2018). Supervised Machine Learning Algorithms for Priority Task Classification in the Cloud Computing Environment. International Journal of Network Security. 18. 176
[21]. H. Zhang and R. Zhou, The analysis and optimization of decision tree based on ID3 algorithm, 2017 9th International Conference on Modelling, Identification and Control (ICMIC), Kunming, China, 2017, pp. 924-928, doi: 10.1109/ICMIC.2017.8321588.
[22]. Gupta, Bhumika & Rawat, Aditya & Jain, Akshay & Arora, Arpit & Dhami, Naresh. (2017). Analysis of Various Decision Tree Algorithms for Classification in Data Mining. International Journal of Computer Applications. 163. 15-19. 10.5120/ijca2017913660
[23]. Nalini Jagtap, P. P. Shevatekar, Nareshkumar Mustary, A comparative study of classification techniques in data mining algorithms, in International Journal of Modern Trends in Engineering and Research, 2017
[24]. Babu, Kunda & Narasimha Rao, Prof. (2023). A Study on Imbalanced Data Classification for Various Applications. Revue D Intelligence Artificielle. 37. 517-524. 10.18280/ria.370229.
[25]. Awad, Mohammed & Fraihat, Salam. (2023). Recursive Feature Elimination with Cross-Validation with Decision Tree: Feature Selection Method for Machine Learning-Based Intrusion Detection Systems. Journal of Sensor and Actuator Networks. 12. 67. 10.3390/jsan12050067.
[26]. Octavian Ceban, Rețele neuronale utilizate în evaluarea performanței pe piața de capital, Colecția de working papers ABC-UL LUMII FINANCIARE, 2018
[27]. M. Danubianu, S.G. Pentiuc, D.M. Danubianu, Data Dimensionality Reduction for Data Mining: A Combined Filter-Wrapper Framework, International Journal of Computers, Communications and Control, ISSN 1841-9836, 2012
[28]. Hernández, Antonio & Herrera-Flores, Boris & Tomás, David & Navarro-Colorado, Borja. (2019). A Systematic Review of Deep Learning Approaches to Educational Data Mining. Complexity. 2019. 1-22. 10.1155/2019/1306039
[29]. Mihaela Chistol, Mirela Danubianu, Survey of Text Mining Research Methods and Their Innovative Applicability, Journal of danubian studies and research, Vol.11, No.1/2021
[30]. Vidhya, K.A.. (2021). Text Mining Process, Techniques and Tools : an Overview. International Journal of Information Technology and Management.
[31]. Tang, Ruixiang & Han, Xiaotian & Jiang, Xiaoqian & Hu, Xia. (2023). Does Synthetic Data Generation of LLMs Help Clinical Text Mining?.
[32]. Ferreira, Rafael & Ferreira, Máverick André & Pinheiro, Anderson & Costa, Evandro & Romero, Cristóbal. (2019). Text mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 9. 10.1002/widm.1332
[33]. Zanini, Nadir & Dhawan, Vikas. (2015). Text Mining: An introduction to theory and some applications. Research Matters. 38-44
[34]. Newman, Heather & Joyner, David. (2018). Sentiment Analysis of Student Evaluations of Teaching. 246-250. 10.1007/978-3-319-93846-2_45
[35]. Ruseti, Stefan & Dascalu, Mihai & Johnson, Amy & Balyan, Renu & Kopp, Kristopher & McNamara, Danielle & Trausan-Matu, Stefan. (2018). Predicting Question Quality Using Recurrent Neural Networks. 10.1007/978-3-319-93843-1_36

[36]. Yoo, Jaebong & Kim, Jihie. (2013). Can Online Discussion Participation Predict Group Project Performance? Investigating the Roles of Linguistic Features and Participation Patterns. International Journal of Artificial Intelligence in Education. 24. 10.1007/s40593-013-0010-8

[37]. Iqbal, Shakeel & Qureshi, Ijaz. (2011). Learning Management Systems (LMS): Inside Matters. Inf. Manag. Bus. Rev.. 3. 206-216. 10.22610/imbr.v3i4.935.

[38]. A. Dutti, MA Ismail și T. Herawan, A Systematic Review on Educational Data Mining, IEEE Access, vol. 5, p. 15991-16005, 2017, doi: 10.1109/ACCESS.2017.2654247

[39]. Muttathil, Anoopkumar & Rahman, A.M.J.Md.Zubair. (2016). A Review on Data Mining techniques and factors used in Educational Data Mining to predict student amelioration. 122-133. 10.1109/SAPIENCE.2016.7684113.

[40]. Şahin, Muhittin & Yurdugül, Halil. (2020). Educational Data Mining and Learning Analytics: Past, Present and Future. 9. 121-131.

[41]. Romero, Cristóbal & Ventura, Sebastian. (2010). Educational Data Mining: A Review of the State of the Art. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on. 40. 601 - 618. 10.1109/TSMCC.2010.2053532.

[42]. Kitchenham, Barbara & Charters, Stuart. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. 2.

[43]. S. Pulakhandam and N. Patil, Recomandation of Optimal Locations for Government Financed Educational Institutes in Urban India Using a Hybrid Data Mining Technique, 2015 Second International Conference on Advances in Computing and Communication Engineering, Dehradun, India, 2015, pp. 560 -567, doi : 10.1109/ICACCE.2015.140.

[44]. https://lege5.ro/Gratuit/heztqmry/legea-nr-1150-1864-asupra-instructiunii-a-principateloru-unite-romane

[45]. D. Marcu and M. Danubianu, Sentiment Analysis from Students' Feedback: A Romanian High School Case Study, 2020 International Conference on Development and Application Systems (DAS), 2020, pp. 204-209, doi: 10.1109/DAS49615.2020.9108927

[46]. C. Simionescu, M. Danubianu, D. Marcu (2020) Analysis of online education romanian schools due to Covid-19 pandemics and areas of improvement, ICERI2020 Proceedings, pp. 3523-3529.

[47]. Simionescu, C., Danubianu, M., Marcu, D., Turcu, CO., Online Learning after One Year of Digital Schooling in Romania: A Survey, International journal of computer science and network security, Vol. 22 No. 1 pp. 27-32 January 2022, https://doi.org/10.22937/IJCSNS.2021.21.12.99

[48]. Joshi, Ankur & Kale, Saket & Chandel, Satish & Pal, Dinesh. (2015). Likert Scale: Explored and Explained. British Journal of Applied Science & Technology. 7. 396-403. 10.9734/BJAST/2015/14975.

[49]. http://analize-statistice.eu/consistenta-interna-cronbach-alfa/

[50]. https://eurydice.eacea.ec.europa.eu/ro/national-education-systems/romania/reforme-derulare-si-evolutii-materie-de-politici

[51]. https://open-science-cloud.ec.europa.eu/

[52]. https://zenodo.org/

[53]. Okpala, I., Romera Rodriguez, G., Tapia, A., Halse, S., & Kropczynski, J. (2022, December). A Semantic Approach to Negation Detection and Word Disambiguation with Natural Language Processing. In Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval (pp. 36-43).

[54]. N. V. Chawla, K. W. Bowyer,L. O. Hall,W. P. Kegelmeyer , SMOTE: Synthetic Minority Over-sampling Technique, Journal Of Artificial Intelligence Research, Volume 16, pages 321-357, 2002, https://doi.org/10.1613/jair.953.

[55]. Wirth, R., & Hipp, J., CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining,Vol. 1, pp. 29-39, 2000

[56]. Ye, Jong Chul. (2022). Generalization Capability of Deep Learning. 10.1007/978-981-16-6046-7_12.

[57]. Taye MM. Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. Computers. 2023; 12(5):91. https://doi.org/10.3390/computers12050091

[58]. Ahmed, Shams & Alam, Md. Sakib & Hassan, Maruf & Rodela, Mahtabin & Ishtiak, Taoseef & Rafa, Nazifa & Mofijur, M. & Ali, A B M Shawkat & Gandomi, Amir. (2023). Deep learning modelling techniques: current progress, applications, advantages, and challenges. Artificial Intelligence Review. 56. 10.1007/s10462-023-10466-8.

[59]. Fu, Erjia & Xiang, Junyuan & Xiong, Chuanhao. (2022). Deep Learning Techniques for Sentiment Analysis. Highlights in Science, Engineering and Technology. 16. 1-7. 10.54097/hset.v16i.2065.

[60]. Ganie, Aadil & Dadvandipour, Samad. (2022). Traditional or deep learning for sentiment analysis: A review. Multidiszciplináris tudományok. 12. 3-12. 10.35925/j.multi.2022.1.1.

[61]. Tang, Duyu & Liu, Ting. (2015). Deep learning for sentiment analysis: Successful approaches and future challenges. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 5. 292-303. 10.1002/widm.1171.