UNIVERSITATEA BABEŞ-BOLYAI
FACULTATEA DE MATEMATICĂ ȘI INFORMATICĂ

Extragerea automată a caracteristicilor comportamentale ale educabililor în diferite medii de învățare

Rezumatul tezei de doctorat

Student doctorand: Mariana-Ioana DINDELEGAN (căs. MAIER)
Conducător științific: prof. dr. Gabriela CZIBULA

2023

Cuvinte cheie: extragerea cunoștințelor din date educaționale, extragerea cunoștințelor din date comportamentale, învățare automată, competențele secolului al XXI-lea, analiza și predicția performanței educabililor.

Cuprins

Cuprinsul tezei de doctorat

Lista publicațiilor

Clasamentul publicațiilor a fost realizat conform standardelor CNATDCU (Consiliul Naţional de Atestare a Titlurilor, Diplomelor si Certificatelor Universitare) aplicabile pentru studenții doctoranzi înscriși după 1 octombrie 2018. Toate clasamentele sunt listate conform clasificării jurnalelor [1] și a conferințelor [2] în Informatică.

Publicații indexate în Web of Science - Science Citation Index Expanded

[MCD23] **Mariana-Ioana Maier**, Gabriela Czibula, Lavinia-Ruth Delean, *Using unsupervised learning for mining behavioural patterns from data. A case study for the baccalaureate exam in Romania*, Studies in Informatics and Control, 32(2), pp. 73-84, 2023. (**AIS Quartila Q4** conform JCR 2022)

Clasificare C, 2 puncte.

[CCML22] Gabriela Czibula, George Ciubotariu, **Mariana-Ioana Maier**, Hannelore Lisei, *IntelliDaM: A machine learning based framework for enhancing the performance of decision-making processes. A case study for educational data mining*, IEEE Access, Volume 10, pp. 80651–80666, 2022 (**AIS Quartila Q2** conform JCR 2021)

Clasificare B, 2 puncte.

[MCOM21] **Mariana-Ioana Maier**, Gabriela Czibula, Zsuzsanna-Edit Oneţ-Marian. *Towards using unsupervised learning for comparing traditional and synchronous online learning in assessing students' academic performance*, Mathematics, Special issue on Didactics and Technology in Mathematical Education, 2021, 9(22), 2870 (**IF Quartila Q1** conform JCR 2020)

Clasificare A, 8 puncte.

[OMCM21] Zsuzsanna Oneţ-Marian, Gabriela Czibula, **Mariana Maier**. *Using self-organizing maps for comparing students' academic performance in online and traditional learning environments* (2021). Studies in Informatics and Control, 30(4), pp. 1–11 (**IF Quartila Q3** conform JCR 2020)

Clasificare C, 2 puncte.

[1] https://uefiscdi.ro/premierea-rezultatelor-cercetarii-articole
[2] http://portal.core.edu.au/conf-Clasificares/

Publicații indexate în Web of Science - Conference Proceedings Citation Index

[McA22] **Mariana Maier**, Camelia Șerban, Andrei Moisin. *Mining Sorting Concept across Curriculum Levels: A Cyclic Learning Based Approach* (2022). The 4th International Workshop on Education through Advanced Software Engineering and Artificial Intelligence, workshop of ESEC/FSE conference, pp. 10-17.

Clasificare A, 6 puncte.

[CCCD20] Liana Maria Crivei, Gabriela Czibula, George Ciubotariu, **Mariana Dindelegan**. *Unsupervised learning based mining of academic data sets for students' performance analysis* (2020). IEEE 14th International Symposium on Applied Computational Intelligence and Informatics, SACI 2020, Timișoara, România, pp. 457-462.

Clasificare D - CORE2020, 0.5 puncte.

Publicații în jurnale și volume ale conferințelor

[Din18] **Mariana Dindelegan** (2018). *Digital and Coding Literacy for School Students*. Studia UBB Digitalia, Volume 63 (LXIII) 2018, June, Issue 1, 55-68. (**indexat Central & Eastern European Online Library**)

Clasificare D, 1 point.

[FPDP+14] Silvia Ferent Pipas, **Mariana Dindelegan**, Bogdan Padurean, Emilia Ciupan, Cornel Ciupan (2014). *Cost calculator for water jet, laser and plasma machining*. Acta Technica Napocensis Series: Applied Mathematics and Mechanics, Volume 57, 2014, March, Issue 1, 73-76. (**indexat Index Copernicus**)

Clasificare D, 0 puncte.

Scorul publicațiilor: 21.5 puncte.

Introducere

Domeniile principale ale prezentei teze de doctorat sunt reprezentate de extragerea cunoștințelor din date educaționale și din date comportamentale. Teza este intitulată „Extragerea automată a caracteristicilor comportamentale ale educabililor în diferite medii de învățare" și își propune să obțină o imagine de ansamblu asupra celor mai importante caracteristici ale elevilor și studenților în diferite medii de învățare. *Extragerea cunoștințelor din date educaționale* (eng. *Educational Data Mining* - EDM) este un domeniu consistent de studiu, în care obiectivul principal este de a identifica tipare semnificative în datele colectate din diferite medii educaționale. *Extragerea cunoștințelor din date comportamentale* (eng. *Behavioural Mining* - BM) este un subdomeniu al *Extragerii cunoștințelor din date* (eng. *Data Mining* - DM) concentrat pe extragerea tiparelor comportamentale din date. Teza noastră își propune să aducă împreună aceste domenii, pentru a obține o imagine a caracteristicilor elevilor și studenților zilelor noastre în diferite medii de învățare.

În mediile educaționale, DM oferă metode pentru oferirea de suport decizional. Descoperirea tiparelor semnificative și extragerea de cunoștințe din seturi de date legate de educație este un subiect provocator și intens investigat în literatura de specialitate EDM [MT13a], în special în urma pandemiei de Covid-19 [GSA21, CcMS+20]. Un obiectiv major în EDM este reprezentat de înțelegerea procesului de învățare al educabililor, predicția rezultatelor învățării, oferirea unei mai bune înțelegeri a fenomenelor legate de educație [BCR18] și sprijinirea instituțiilor de învățământ în înțelegerea și îmbunătățirea proceselor legate de educație. În prezent, instituțiile academice sunt din ce în ce mai interesate să-și îmbunătățească metodologiile de predare, procesele de învățare [MT13b] și performanța academică a studenților și profesorilor [JRHR15]. EDM abordează tehnici de înțelegere a proceselor de învățare și de identificare a tiparelor în date, pentru sprijinirea instituțiilor academice în luarea deciziilor (cu privire la admiterea la universitate [Men20] sau influențarea performanței studenților în anii lor de facultate [AS19]).

Fiecare furnizor educațional și, în general, fiecare furnizor de servicii, încearcă să ofere produse adecvate beneficiarilor săi. În acest sens, furnizorii trebuie să aibă o imagine adecvată a performanței clienților, astfel încât produsele sau serviciile oferite să poată fi adaptate în funcție de aceste performanțe.

Având în vedere evoluția rapidă a societății, se impune o schimbare de paradigmă în educație. Prin urmare, sistemele de învățământ trebuie să ia în considerare instrumentele disponibile, astfel încât această schimbare să poată aduce beneficii educabililor, profesorilor și instituțiilor de învățământ. De exemplu, odată cu criza Covid-19, educația s-a orientat către *medii online*, iar metodele tradiționale de predare utilizate de instituțiile de învățământ au trebuit să fie adaptate. S-a constatat că succesul în învățare (atât în contexte tradiționale, cât și online) este influențat de motivația elevilor și de eficacitatea profesorilor. Calitatea predării nu garantează motivaţia elevilor sau invers, deoarece aceasta din urmă depinde de alţi factori, intrinseci sau extrinseci [Nas20]. În acest context, există un interes din ce în ce mai mare pentru înțelegerea modului în care studenții învață și cum ar putea să-și îmbunătățească performanța academică.

Problema abordată

Principala problemă abordată în prezenta teză este extragerea caracteristicilor comportamentale ale elevilor și studenților, pentru că înțelegerea modului în care educabilii își construiesc și îmbunătățesc rezultatele este un subiect de interes tot mai mare pentru fiecare furnizor de educație.

Prima noastră preocupare este să identificăm *caracteristici comportamentale la elevi*, pentru a-i ajuta pe profesori în procesul instructiv-educativ. Pe parcursul carierei didactice (aproximativ 40 de ani), se perindă diferite generații de elevi, deci profesorii sunt nevoiți să își adapteze în permanență metodele didactice ținând cont de profilul generației curente, pentru că fiecare vârstă are motivația ei, interesele ei, abilitățile ei etc. iar elevii trebuie implicați activ în învățare. Identificarea caracteristicilor comportamentale ale elevilor ne-ar putea da profilul generației actuale de elevi, cu interesele și abilitățile ei și, astfel, i-am putea ajuta pe profesori să-și adapteze metodele didactice în funcție de elevii la care predau.

Explorarea conceptelor din informatică sau din alte discipline școlare ar putea ajuta instituțiile școlare în proiectarea și adaptarea curriculei la generația actuală de elevi, astfel încât să le faciliteze procesul de învățare. Elevii de azi aparțin Generației Z (cei născuți între 1995 și 2010 [SG16], numiți și „nativi digitali"). Cei din Generația Z sunt mai avansați din punct de vedere tehnologic și, probabil, mai independenți decât cei din generațiile anterioare [MCS19]. Totuși, deși sunt numiți „nativi digitali", ei nu sunt pregătiți să utilizeze inovațiile tehnologice în scopuri tactice sau să-și construiască o carieră. Potrivit lui Shatto și Erwin [Sha16], cei din Generația Z se conectează ușor la serviciile media, ceea ce le permite să studieze un subiect oricând și oriunde doresc, pe diferite dispozitive. Dependența lor de tehnologie are impact direct asupra capacității lor de a învăța. Dacă explorăm conceptele necesare pentru societatea noastră și ținem cont de profilul Generației Z, putem îmbunătăți proiectarea instrucțională, începând cu nivelul școlar și continuând cu cel universitar.

A doua preocupare este legată de *extragerea caracteristicilor comportamentale ale studenților*. La acest nivel, dorim să identificăm tendințe în timpul evaluării performanței studenților atât în mediul de învățare tradițional, cât și în mediul online. Intenționăm să utilizăm tehnici de *învățare nesupervizată* (eng. *Unsupervised Learning* - UL), pentru analiza și predicția performanței studenților și pentru a compara analiza rezultatelor obținute din mediul de învățare tradițional cu cele din mediul online. În contextul prezent, când unele activități de predare și evaluare, inclusiv cursurile, temele și examenele, sunt mutate în medii online, există un interes crescut pentru înțelegerea procesului de învățare al studenților, cu scopul de a le îmbunătăți rezultatele.

În consecință, există o dezvoltare accelerată a acestor aspecte. Epidemia de Covid-19, de exemplu, a modificat fiecare aspect al vieții de zi cu zi, inclusiv școlarizarea. *Învățarea online* a fost un potențial remediu pentru metodele tradiționale de predare ale furnizorilor de educație. Eficacitatea învățării online depinde de standardele de instruire și de implicarea studenților. Aceste aspecte ne motivează să facem o analiză comparativă a performanței studenților obținute în mediul tradițional cu cea obținută în mediile online.

A treia preocupare este *să dezvoltăm o metodologie bazată pe învățarea automată pentru analiza performanței studenților*. Un asemenea instrument ar putea ajuta instituțiile educaționale în procesul decizional. În acest sens, dorim să introducem și să validăm metodologia noastră pe seturi de date educaționale colectate din mediul universitar.

În demersul nostru, provocarea este dată de două aspecte majore: (1) insuficiența seturilor de date cu acces deschis (am fost nevoiți să ne colectăm seturile de date și am observat că oamenii au rezerve atunci când trebuie să împărtășească informații personale, chiar dacă răspunsurile sunt anonime) și (2) dificultatea de a realiza comparații cu literatura de specialitate (sistemele educaționale au particularități esențiale, în funcție de politica țării lor).

Contribuții originale

Cercetarea este orientată pe trei subiecte majore: (1) investigarea trăsăturilor comportamentale ale elevilor; (2) explorarea caracteristicilor comportamentale ale studenților; și (3) propunerea unei metodologii bazate pe învățare automată pentru analiza performanței studenților. Pentru aceste direcții, ne-am concentrat pe: metode de *analiză statistică* (testul Chi-Sqare, testul *Wilcoxon signed-rank*, coeficienții de corelație Pearson și Spearman), metode de *învățare nesupervizată* (partiționare *k-means*, analiza componentei principale (eng. *principal component analysis* - PCA), hărți de auto-organizare (eng. *self-organizing maps* - SOMs), sisteme cu autosupervizare (eng. *autoencoders* - AEs), *t-distributed stochastic neighbor embedding* (t-SNE), aproximare și proiecție uniformă a varietății (eng. *Uniform Manifold Approximation and Projection* - UMAP) și reguli de asociere (eng. *association rules* - ARs)), metode de *învățare supervizată* (regresie logistică, regresie liniară, analiză discriminantă liniară (eng. *Liniar Discriminant Analysis* - LDA), regresie polinomială, gradientul stocastic descendent (eng. *Stochastic gradient descent* - SGD), regresie Tweedie) și *selectarea caracteristicilor* (algoritmul ReliefF).

Astfel, rezultatele și contribuțiile noastre principale sunt, de asemenea, separate în aceste trei direcții, prezentate în capitolele 2, 3 și 4:

1. **Caracteristicile comportamentale ale elevilor**

   Am început cercetarea noastră în mediul școlar, investigând trăsăturile comportamentale ale elevilor cu vârsta cuprinsă între 10 și 19 ani. În acest sens, am propus o analiză statistică pentru a extrage perspective semnificative, *k-means* și SOMs pentru partiționare și *reguli de asociere* pentru a identifica relațiile și dependențele interesante dintre variabilele prezente în seturile noastre de date. Pentru această direcție, seturile de date au fost obținute prin aplicarea chestionarelor subiecților vizați. Rezultatele pe această linie de cercetare au fost următoarele:

   (a) Primul experiment a avut ca scop identificarea competențelor digitale la elevii de clasa a IV-a, ca trăsături comportamentale legate de competențele necesare secolului al XXI-lea. Metodologia și rezultatele experimentale sunt detaliate în secțiunea a doua din Capitolul 2. Detaliile și rezultatele muncii noastre în acest sens au fost publicate în articolul „Digital and Coding Literacy for School Students” [Din18].

   (b) Pentru a explora trăsăturile comportamentale la elevii de la un alt nivel de studiu, ne-am concentrat pe elevii de liceu, cu scopul de a identifica preferințele acestora în alegerea probei opționale la examenul de bacalaureat. Rezultatele sunt publicate în articolul „Using Unsupervised Learning for Mining Behavioural Patterns from Data. A Case Study for the Baccalaureate Exam in Romania” [MCD23]. Metodologia și rezultatele sunt detaliate în a treia secțiune din Capitolul 2.

   (c) O altă perspectivă a fost explorarea conceptelor din informatică la diferite niveluri curriculare, folosind o abordare bazată pe învățarea ciclică. Am început cu conceptul de sortare, în scopul de a conecta cunoștințele și competențele dobândite în gimnaziu și liceu cu cerințele de la nivelul academic. Studiul a fost prezentat și publicat cu titlul „Mining sorting concept across curriculum levels: a cyclic learning based approach” [McA22]. Metodologia și rezultatele sunt detaliate în ultima secțiune a Capitolului 2.

2. **Caracteristicile comportamentale ale studenților**

   A doua direcție pentru prezenta teză a fost să analizăm caracteristicile comportamentale ale studenților. Am folosit analiza statistică pentru a observa corelații semnificative, metode de

învățare nesupervizate pentru analiza performanței studenților și metode de *învățare supervizată* pentru a consolida rezultatele obținute în abordarea *învățării nesupervizate*. Pentru această direcție, seturile de date utilizate în cercetare au fost obținute din activitatea studenților de la Universitatea Babeş-Bolyai Cluj-Napoca, Facultatea de Matematică și Informatică la discipline de specialitate. Rezultatele noastre pe această linie de cercetare au fost următoarele:

(a) În primul rând, ne-am dorit să extragem cunoștințe din seturi de date obținute la disciplina „Structuri de date și algoritmi", analizând performanța studenților prin metode de învățare nesupervizată. Ideea urmărită a fost extragerea de RARs și identificarea partițiilor din seturile de date care conțineau notele obținute de studenți pe parcursul unui semestru. Studiul a fost prezentat și publicat cu titlul „Unsupervised learning based mining of academic data sets for students' performance analysis" [CCCD20], iar metodologia și rezultatele obținute sunt prezentate în prima secțiune a Capitolului 3.

(b) Evaluarea comparativă a performanței academice în mediile de învățare tradiționale și cele online a fost un alt obiectiv al cercetării noastre, ca o consecință a pandemiei de Covid-19 și a trecerii activităților de instruire în mediul online. A doua secțiune din Capitolul 3 prezintă rezultatele obținute în această direcție. Rezultatele au fost publicate în două studii despre performanța studenților la cursul „Programare logică și funcțională", și anume:

   i. Utilizarea învățării nesupervizate pentru a compara învățarea tradițională și online sincron în evaluarea performanței academice a studenților [MCOM21].

   ii. Utilizarea *hărților de auto-organizare* pentru a compara performanțele academice ale studenților în mediile de învățare online și tradiționale [OMCM21].

3. **O metodologie bazată pe învățarea automată pentru analiza performanței studenților**

   A treia direcție în cercetarea noastră a fost construirea unei metodologii pentru analiza performanței studenților. Aceasta a fostă numită $IntelliDaM$ și are trei componente principale: (1) analiza și selecția caracteristicilor; (2) analiza datelor bazată pe *învățare nesupervizată*; și (3) modele predictive bazate pe *învățare supervizată*. Pentru evaluarea performanței $IntelliDaM$, am folosit date obținute de la Universitatea Babeş-Bolyai, România, pe parcursul a trei ani academici, pentru o disciplină informatică. Studiul este publicat cu titlul „$IntelliDaM$: A Machine Learning-Based Framework for Enhancing the Performance of Decision-Making Processes. A Case Study for Educational Data Mining" [CCML22]. Metodologia și rezultatele sunt detaliate în capitolul 4.

Structura tezei

Teza este structurată după cum urmează.

Capitolul 1 prezintă fundamentul teoretic și literatura de specialitate. Se începe cu prezentarea domeniului EDM, cu analiza și predicția performanței educabililor, iar apoi se evidențiază BM aplicată în domeniul educațional. Al doilea aspect din acest capitol este analiza profilului elevilor în contextul digitalizării, pentru a valida competențele necesare secolului al XXI-lea, alfabetizarea digitală din prezent, gamificarea ca strategie educațională, taxonomiile de învățare utilizate în educația actuală, precum și învățarea online și digitalizarea introduse în scopul schimbării de paradigmă în educație. Cea de-a treia secțiune tratează modelele de învățare automată utilizate în teza noastră: modele de învățare nesupervizată (partiționarea *k-means*, PCA, t-SNE, UMAP, SOMs, AEs și ARs), modele de învățare supervizată utilizate pentru a valida rezultatele analizei nesupervizate (LDA, regresie liniară, regresie polinomială, regresie logistică, SGD și regresorul Tweedie) și selecția caracteristicilor.

În Capitolul 2, prezentăm rezultatele noastre din extragerea caracteristicilor comportamentale ale elevilor în procesul de învățare. Prima secțiune prezintă metodele de analiză statistică utilizate în experimentele noastre. A doua secțiunea exploatează competențele necesare secolului al XXI-lea la elevii de clasa a IV-a din cinci școli din România în anul 2018. Metodele utilizate sunt analiza statistică și partiționarea *k-means*. În a treia secțiune, ne propunem să identificăm câteva tendințe ale elevilor de liceu în alegerea disicplinei opționale la examenul de bacaluareat prin metode de învățare nesupervizată (SOMs și ARs) și analiză statistică. Ultima secțiune este o incursiune în curriculumul românesc de informatică, cu scopul de a observa abordarea conceptului de sortare la mai multe niveluri de cunoaștere și de a propune o metodologie utilă educabililor în învățarea algoritmilor de sortare. Ne-au interesat nivelurile școlare de gimnaziu și liceu și impactul lor la nivel de facultate.

În Capitolul 3 este prezentată munca noastră legată de extragerea caracteristicilor de învățare ale studenților. În prima secțiune am folosit RARs și PCA pentru a analiza performanța studenților la cursul „Structuri de date și algoritmi". A doua secțiune prezintă analiza comparativă a performanței studenților în mediul tradițional versus mediul online la cursul „Programare logică și funcțională". În acest sens, am folosit tehnici de învățare nesupervizată pentru analiză, cum ar fi: AEs, t-SNE, PCA și SOMs și unele tehnici de învățare supervizată, precum regresia logistică, regresia liniară și LDA pentru a valida analiza noastră.

Capitolul 4 descrie metodologia propusă de noi, numită $IntelliDaM$, pentru analiza performanței studenților. Pe lângă metodologia propusă, contribuțiile suplimentare avute în vedere de cercetarea noastră sunt: (1) evidențierea eficienței metodologiei $IntelliDaM$ în analiza datelor legate de performanța studenților; (2) analiza și interpretarea, pentru studiul de caz considerat, a relevanței tiparelor extrase în mod nesupervizat din datele academice și a modului în care aceste tipare sunt corelate cu performanța academică a studenților; și (3) verificarea faptului că predicția finală a performanței studenților la o anumită disciplină academică este sau nu îmbunătățită de rezultatele obținute la cursurile anterioare de informatică din curriculum.

După descrierea muncii noastre în aceste capitole, sunt prezentate concluziile cu extinderile ulterioare, apoi anexele și lista resurselor bibliografice.

Capitolul 1

Fundamente teoretice

Acest capitol descrie, pe scurt, conceptele de bază pentru domeniul tezei noastre, prezintă lucrările aferente din literatura de specialitate și principalele modele computaționale pe care le folosim în cercetare. Capitolul este organizat după cum urmează.

Prima secțiune prezintă concepte utile din EDM, și anume *Analiza și predicția performanței educabililor* și *Extragerea cunoștințelor din date comportamentale în domeniul educațional*.

Secțiunea a doua introduce problema analizei profilului educabililor în societatea noastră digitalizată. Secțiunea începe cu o scurtă incursiune în *competențele necesare secolului al XXI-lea*, apoi este prezentat conceptul de *alfabetizare digitală* din societatea noastră. *Gamificarea* este descrisă apoi, pentru că acest concept este din ce în ce mai utilizat în procesul instrucțional al zilelor noastre. În continuare, sunt prezentate principalele *taxonomii ale învățării*, cu accent pe *Taxonomia Revizuită a lui Bloom*. La sfârșitul secțiunii, ne concentrăm pe *instruire online și digitalizare*.

Ultima secțiune descrie modelele de *învățare automată* (eng. *Machine Learning* - ML) utilizate în studiile noastre. Secțiunea este alcătuită din trei părți: *învățare nesupervizată* (partiționare *k-means*, analiza componentei principale, *t-distributed stochastic neighbor embedding*, aproximare și proiecție uniformă a varietății, hărți auto-organizate, sisteme cu autosupervizare și reguli de asociere), *învățare supervizată* (analiza discriminantă liniară, regresia liniară, regresia polinomială, regresia logistică, gradientul stocastic descendent și regresia Tweedie) și *selectarea caracteristicilor* împreună cu algoritmii reprezentativi.

Capitolul 2

Explorarea caracteristicilor comportamentale ale elevilor

Acest capitol este o incursiune în câmpul caracteristilor comportamentale ale elevilor, cu scopul de a schița profilul de învățare al elevilor contemporani. Capitolul este structurat după cum urmează.

Prima secțiune prezintă concepte din analiza statistică utilizate în studiile noastre: testul *Chi-Square*, *Z-score* și *F-value*.

În a doua secțiune, dorim să identificăm *competențele necesare secolului al XXI-lea* în profilul elevilor de clasa a IV-a, în cadrul proiectului educațional al Asociației „Cartea Daliei". Rezultatele obținute sunt publicate în articolul cu titlul „Digital and Coding Literacy for School Students" [Din18]. Scopul studiului a fost, pe de o parte, să determinăm în ce măsură sunt pregătiți elevii de clasa a IV-a să înceapă orele obligatorii de Informatică și TIC de la nivel gimnazial și, pe de altă parte, să schițăm profilul de învățare al elevilor, identificându-le competențele necesare pentru provocările societății noastre digitalizate.

Secțiunea a treia continuă schițarea *profilului elevilor* la nivel liceal, în scopul obținerii celor mai importante caracteristici care îi influențează în alegerea disciplinei opționale la examenul de bacalaureat din România. Această secțiune prezintă studiul desfășurat pentru elevii claselor cu profil real, adică cei de la specializările: *Matematică-Informatică*, *Matematică-Informatică intensiv Informatică* și *Științe ale Naturii*. Studiul este publicat sub titlul „Using Unsupervised Learning for Mining Behavioural Patterns from Data. A Case Study for the Baccalaureate Exam in Romania" [MCD23]. Una din concluziile desprinse este că elevii de liceu sunt foarte interesați de alegerea disciplinelor Biologie și Informatică la examenul de bacalaureat.

După ce am observat interesul mare al elevilor de liceu pentru Informatică în secțiunea a treia, ne-am propus să realizăm studii transversale pe parcursul nivelurilor curriculare pentru cele mai importante concepte din Informatică. Ultima secțiune a acestui capitol prezintă munca noastră cu privire la abordarea *conceptului de sortare pe parcursul a trei niveluri curriculare*: gimnaziu, liceu și facultate, cu accent pe primele două niveluri. Acest studiu urmărește perspectiva învățării ciclice și a fost prezentat și publicat sub titlul „Mining sorting concept across curriculum levels: a cyclic learning based approach" [McA22]. Ca extindere, propunem o aplicație care să îi ajute pe elevi și profesori în procesul instrucțional al algoritmilor de sortare. Această aplicație va colecta date comportamentale ale utilizatorilor în timpul învățării algoritmilor de sortare, iar datele colectate vor fi utilizate în crearea profilului elevilor care vor folosi aplicația.

Capitolul 3

Explorarea caracteristicilor comportamentale ale studenților

Capitolul 3 continuă procesul de explorare a caracteristicilor comportamentale pentru educabilii de nivel academic. Pentru această direcție, ne concentrăm pe predicția și analiza performanței studenților utilizând metode bazate pe *învățare nesupervizată* (reguli relaționale de asociere, analiza componentei principale, sisteme cu autosupervizare, t-SNEs și hărți auto-organizate). Câteva metode bazate pe *învățarea supervizată* (regresia liniară, regresia logistică, analiza discriminantă liniară) au fost utilizate pentru validarea rezultatelor obținute prin intermediul tehnicilor bazate pe învățare nesupervizată. Au fost utilizate seturi de date reale, colectate de la Facultatea de Matematică și Informatică a Universității Babeș-Bolyai din România.

Prima secțiune prezintă utilizarea *regulilor relaționale de asociere* și a *analizei componentei principale* în analiza performanței studenților la cursul *Structuri de date și algoritmi*. Studiul este publicat sub titlul „Unsupervised learning based mining of academic data sets for students' performance analysis" [CCCD20].

A doua secțiune a fost inspirată de trecerea activităților educaționale în mediul online, ca o consecință a pandemiei de Covid-19. În această secțiune desfășurăm două analize comparative cu privire la performanțele obținute de studenți în mediul tradițional versus performanțele obținute în mediile de învățare online pentru cursul *Programare logică și funcțională*. Aceste studii au fost diseminate prin articolele:

- "Towards using unsupervised learning for comparing traditional and synchronous online learning in assessing students' academic performance" [MCOM21].
- "Using Self-Organizing Maps for Comparing Students' Academic Performance in Online and Traditional Learning Environments" [OMCM21].

Capitolul 4

$IntelliDaM$: O metodologie bazată pe învățarea automată pentru analiza performanței studenților

În scopul de a îmbunătăți rezultatele analizei performanței studenților, în Capitolul 4 am introdus $IntelliDaM$, o metodologie bazată pe învățarea automată pentru extragerea cunoștințelor din datele legate de performanța studenților. $IntelliDaM$ oferă trei componente de analiză a datelor, concepute pentru: (1) analiza și selecția caracteristicilor; (2) analiza datelor bazată pe *învățare nesupervizată*; și (3) modele predictive bazate pe *învățare supervizată*.

Pentru evaluarea performanței metodologiei $IntelliDaM$, am folosit seturi de date reale, obținute de la Universitatea Babeș-Bolyai, România, pe parcursul a trei ani academici la disciplina *Programare logică și funcțională*. Pe lângă metodologia propusă, contribuțiile suplimentare preconizate de cercetarea noastră au fost: (1) sublinierea eficacității metodologiei $IntelliDaM$ în analizarea datelor legate de performanța studenților; (2) analizarea și interpretarea, pentru studiul de caz discutat, a relevanței tiparelor extrase în mod nesupervizat din date academice și a modului în care aceste tipare sunt corelate cu performanța studenților; și (3) verificarea faptului că predicția perfomanței finale a studenților la o diciplină academică poate fi îmbunătățită de rezultatele studenților obținute la cursuri anterioare de informatică din curriculum. Chiar dacă este evaluată empiric pe date academice, metodologia propusă $IntelliDaM$ este una generală și poate fi aplicată pentru orice sarcină de analiză a datelor.

Capitolul 4 este structurat după cum urmează. Prima secțiune discută metodele utilizate pentru construirea $IntelliDaM$ și îi prezintă componentele principale. A doua secțiune descrie evaluarea experiențială a metodologiei propuse de noi pe un studiu de caz din domeniul EDM, cu accent pe setul de date, experimentele și rezultatele obținute, în timp ce secțiunea a treia discută rezultatele obținute. Ultima secțiune este reprezentată de concluziile studiului și idei pentru extinderi ulterioare.

Rezultatele obținute în urma acestui studiu sunt prezentate în articolul intitulat „$IntelliDaM$: A Machine Learning-Based Framework for Enhancing the Performance of Decision-Making Processes. A Case Study for Educational Data Mining” [CCML22].

Concluzii

Prezentul document descrie rezultatele originale obținute pentru teza noastră de doctorat intitulată „*Extragerea automată a caracteristicilor comportamentale ale educabililor în diferite medii de învățare*" cu scopul de a dezvolta și implementa tehnici DM în probleme ce aparțin domeniului educațional. Aplicarea tehnicilor DM în educație [BCR18] este, în zilele noastre, un domeniu de cercetare interesant și atractiv, în care scopul principal este construirea de metode de obținere a informațiilor relevante din datele educaționale pentru a înțelege mai profund procesele de învățare și pentru a oferi perspective suplimentare în domeniul educațional. Teza descrie activitatea noastră în acest domeniu.

Am prezentat problemele care ne-au preocupat și contribuțiile noastre originale de până acum. De asemenea, am subliniat direcțiile în care dorim să ne continuăm cercetarea în domeniul EDM.

Prima direcție pe care am vizat-o a fost utilizarea metodelor de învățare nesupervizată și statistică pentru analiza comportamentului elevilor din zilele noastre, pentru a afla care sunt abilitățile, punctele slabe sau factorii de influență ale acestora în procesul lor de instruire. Dorim să corelăm aceste rezultate cu situația actuală din sistemul educațional, pentru a ajuta instructorii să-și adapteze activitatea educațională pentru prezent.

A doua direcție prezentată este utilizarea modelelor de învățare automată la studenți, pentru predicția și analiza performanței studenților. O preocupare importantă a fost compararea performanței studenților în mediile de învățare online și tradiționale, deoarece una dintre marile provocări din ultimii ani a fost trecerea dinspre învățarea tradițională spre cea online.

A treia direcție a fost crearea unei metodologii bazate pe învățarea automată pentru analiza performanței studenților, numită $IntelliDaM$. Această metodologie constă în componente pentru analiza caracteristicilor, tehnici de învățare nesupervizată și tehnici de învățare supervizată, utile în îmbunătățirea performanței sarcinilor de extragere a cunoștințelor din date. Împreună cu primele două direcții, această direcție poate servi profesorilor în alegerea celor mai bune metode și instrumente pentru a face față situației prezente din educație.

Considerăm această abordare foarte importantă pentru situația actuală din domeniul educațional și suntem încrezători că va oferi răspunsuri valoroase pentru oricine este interesat de EDM.

Lucrările viitoare ne vor extinde cercetarea prin dezvoltarea de noi modele ML cu scopul de a descoperi alte tipare semnificative în seturile noastre de date. Un alt scop este să ne îmbogățim metodele de colectare a datelor și astfel, să ne creștem resursele cu date de la educabili aflați în diferite stadii de învățare. De asemenea, intenționăm să luăm în considerare mai multe instrumente de cercetare și să cooperăm cu specialiști din psihologie, pedagogie sau sociologie care ar putea să-și ofere expertiza în cadrul domeniului nostru de cercetare.

Bibliografie

[AS19] A. I. Adekitan and O. Salau. The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon*, 5(2):e01250, 2019.

[BCR18] A. Bogarín, R. Cerezo, and C. Romero. A survey on educational process mining. *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery*, 8(1), 2018.

[CCCD20] L. M. Crivei, G. Czibula, G. Ciubotariu, and M. Dindelegan. Unsupervised learning based mining of academic data sets for students' performance analysis. In *IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI 2020)*, pages 11–16. IEEE Hungary Section, 2020.

[CCML22] Gabriela Czibula, George Ciubotariu, Mariana-Ioana Maier, and Hannelore-Inge Lisei. $IntelliDaM$: A machine learning based framework for enhancing the performance of decision-making processes. A case study for educational data mining. *IEEE Access*, 10:80651–80666, 2022.

[CcMS+20] C. Coman, L. G. Ţîru, L. Meseşan-Schmitz, C. Stanciu, and M. C. Bularca. Online teaching and learning in higher education during the coronavirus pandemic: Students' perspective. *Sustainability*, 12(24), 2020.

[Din18] Mariana Dindelegan. Digital and coding literacy for school students. *Studia UBB Digitalia*, 63(1):55–68, 2018.

[FPDP+14] S. Ferenţ Pipaş, M. Dindelegan, B. Pădurean, E. Ciupan, and C. Ciupan. Cost calculator for water jet, laser and plasma machining. *ACTA TECHNICA NAPOCENSIS Series: Applied Mathematics and Mechanics*, 57:73–76, 3 2014.

[GSA21] Ram Gopal, Varsha Singh, and Arun Aggarwal. Impact of online classes on the satisfaction and performance of students during the pandemic period of covid 19. *Education and information technologies*, pages 1–25, 2021.

[JRHR15] Syed Tanveer Jishan, Raisul Islam Rashu, Naheena Haque, and Rashedur M. Rahman. Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. *Decision Analytics*, 2(1):1, Mar 2015.

[McA22] Mariana Maier, Camelia Şerban, and Moisin Andrei. Mining sorting concept across curriculum levels. a cyclic learning based approach. pages 10–17, 2022.

[MCD23] M. Maier, G. Czibula, and L. Delean. Using unsupervised learning for mining behavioural patterns from data. a case study for the baccalaureate exam in romania. *Studies in Informatics and Control*, 2023.

[MCOM21] Mariana-Ioana Maier, Gabriela Czibula, and Zsuzsanna-Edit Onet-Marian. Towards using unsupervised learning for comparing traditional and synchronous online learning in assessing students' academic performance. *Mathematics, Engineering Mathematics - special issue on Didactics and Technology in Mathematical Education*, 9(22):2870, 2021.

[MCS19] J. B. Mosca, K. P. Curtis, and P. G. Savoth. New approaches to learning for generation Z. *Journal of Business Diversity*, 19(3), 2019.

[Men20] Hanan Abdullah Mengash. Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access*, 8:55462–55470, 2020.

[MT13a] Siti Khadijah Mohamad and Zaidatun Tasir. Educational data mining: A review. *Procedia - Social and Behavioral Sciences*, 97(Supplement C):320 – 324, 2013. The 9th International Conference on Cognitive Science.

[MT13b] Siti Khadijah Mohamad and Zaidatun Tasir. Educational data mining: A review. *Procedia - Social and Behavioral Sciences*, 97:320 – 324, 2013. The 9th International Conference on Cognitive Science.

[Nas20] S. minda Nasution. Online-Learning and Students' Motivation: A Research Study on the Effect of Online Learning on students' motivation in IAIN Padangsidimpuan. *Asian Social Science and Humanities Research Journal (ASHREJ)*, 2(2):9–16, 2020.

[OMCM21] Z. Onet-Marian, G. Czibula, and M. Maier. Using self-organizing maps for comparing students' academic performance in online and traditional learning environment. *Studies in Informatics and Control*, 30(4):1–11, 2021.

[SG16] Corey Seemiller and Meghan Grace. *Generation Z Goes to College*. Jossey-Bass; 1st edition, 2016.

[Sha16] K. Shatto, B.and Erwin. Moving on from millennials: Preparing for generation Z. *Journal of Continuing Education in Nursing*, 47(6):253–254, 2016.