1. **Introducere**- din fiecare articol parcurs ar trebui să extrageți scopul, abordarea, concluziile și probleme rămase nerezolvate. Puteți structura articolele pe categorii.

[1]

Scop:

Scopul acestei lucrări este de a propune un framework mai flexibil pentru a prezice performanța academică a studenților. În acest cadru, datele brute sunt utilizate direct pentru a construi modelul de predicție fără pasul de feature engineering.

Abordare:

Renunțând la pasul de feature engineering, studiul se axează pe selecția caracteristicilor bazată pe interpretabilitatea modelului. Framework-ul este aplicat setului de date Open University Learning Analytics Dataset (OULAD) folosind două tipuri diferite de clasificatoare: Random Forest și Rețele Neuronale Artificiale (RNA).

Acest studiu detaliază trei contribuții principale pentru îmbunătățirea flexibilității și automatizării modelelor de predicție:

1) se propune un cadru care reduce complexitatea abordărilor disponibile în literatură prin eliminarea etapei de inginerie a caracteristicilor;

2) sunt introduse metode de interpretare a modelului pentru a automatiza selecția caracteristicilor pentru diferite scenarii de predicție;

3) cadrul propus este verificat cu două modele populare de predicție disponibile în literatură: RNA și Random Forest.

Această lucrare utilizează numeroase înregistrări ale unor cursanți, ce sunt disponibile în setul de date OULAD. Aceste înregistrări conțin datele demografice ale cursanților și interacțiunile acestora cu un mediu de învățare virtual. OULAD este selectat, de asemenea, deoarece un număr suficient de rezultate au fost publicate folosind acest set de date până acum.

Caracteristicile memorate în setul de date OULAD pot fi clasificate în trei tipuri: informațiile și datele demografice ale studenților, informațiile despre cursuri și interacțiunile cu mediul virtual de învățare. Toate caracteristicile însumează un total de 32.

Ținta predicțiilor din această lucrare este performanța academică a studenților. În OULAD, această țintă este exprimată prin patru categorii diferite: Distincție, Reușită, Eșuare și Retragere. Repartizarea celor 32.593 de studenți pe baza rezultatului final este următoarea: Distincție (3.024), Reușit (12.361), Nereușit (7.052) și Retras (10.156).

Pentru simplitate, se decide să se împartă cronologia cursului în patru momente de predicție distribuite uniform pe durata cursului. Durata cursurilor variază între 234 și 269 de zile. Deci, fiecare curs are momente diferite de predicții. Primul trimestru conține toate interacțiunile studenților până în ultima zi a trimestrului. Aceste interacțiuni alături de caracteristicile statice sunt utilizate pentru a construi modelul de predicție. Datele din trimestrul al doilea adaugă interacțiunile nou înregistrate până la jumătatea cursului. Același raționament este aplicat pentru trimestrul al treilea și al patrulea. Deoarece nimeni nu este interesat să aibă momente de predicție în ultima zi a cursului, sfârșitul celui de-al patrulea trimestru este mutat cu două săptămâni înainte de ultima zi a cursului. Cu alte cuvinte, trimestrul al patrulea conține interacțiunile studentului înregistrate disponibile din primul trimestru, al doilea trimestru, al treilea trimestru și interacțiunile înregistrate până cu 14 zile înainte de ultima zi a cursului.

Această secțiune raționalizează motivul pentru eliminarea etapei de inginerie a caracteristicilor din cadrul propus, urmată de o descriere a două tipuri diferite de modele de predicție utilizate în experimente. Alegerea acestor modele de predicție se bazează pe două criterii: 1) utilizarea lor frecventă în literatură; 2) faptul că provin din două familii distincte de algoritmi de învățare automată. În cele din urmă, este explicată posibilitatea de a lega un pas flexibil de selectare/reducere a caracteristicilor la modelul propus.

From CHPATER 4.1 ALMOST ALL

Concluzii:

Rezultatele obținute arată că etapa de inginerie a caracteristicilor poate fi abandonată fără a afecta performanța de predicție a modelelor. Rezultatele de predicție ale cadrului flexibil de selecție a caracteristicilor fie depășesc, fie au o diferență de precizie mai mică de 1% în comparație cu alte lucrări din literatură care se bazează pe un pas manual de inginerie a caracteristicilor.

Atât rețelele de pădure aleatoare, cât și rețelele neuronale artificiale fără inginerie de caracteristici realizează o precizie de predicție ridicată pentru cazul studenților cu risc de eșec cu 86% și 88% comparativ cu toți studenții cu note de promovare și, respectiv, elevii cu note de distincție.

De asemenea, modelele de predicție au cea mai mare rată de acuratețe de 93% în prezicerea studenților care abandonează școala.

Cu toate acestea, modelele de predicție din cadrul propus și lucrările anterioare de cercetare au rezultate slabe în prezicerea studenților cu rezultate înalte, cu o acuratețe maximă de 81%, o precizie de 69% și un recall de 57%.

Fiecare model a fost investigat în raport cu 16 momente de predicție, patru sferturi din fiecare dintre cele patru probleme de clasificare binară.

Rezultatele obținute din modelele de predicție arată trei concluzii importante:

1) pașii de inginerie a caracteristicilor nu contribuie semnificativ la acuratețea modelelor de predicție, astfel încât aceștia pot fi eliminați;

2) un clasificator Random Forest este la fel de competent ca un clasificator de Deep Learning pentru predicția performanței școlare a studenților;

Modelele de predicție din cadrul propus sunt evaluate în trimestrul al patrulea cu o acuratețe de 85,83%, 88,15%, 81,26% și 92,91%, pentru Pass-Fail, Distinction-Fail, Distinction-Pass și, respectiv, Retras-Pass.

Aceste rezultate fie depășesc, fie au o diferență de mai puțin de 1% în comparație cu precizia de ultimă generație, care se bazează pe un pas manual de inginerie a caracteristicilor.

Studiul nostru arată, de asemenea, că atât cadrul propus, cât și literatura disponibilă au rezultate slabe în clasificarea elevilor cu note de distincție față de elevii cu note de promovare.

Probleme ramase nerezolvate:

Lucrările viitoare se vor concentra pe dezvoltarea unui algoritm/cadru care poate face față unui set de date mic.

Mai mult, informațiile demografice ale studenților ajută la prezicerea performanței acestora, dar utilizarea lor greșită - de exemplu, utilizarea lor pentru profilare - poate duce la consecințe negative neașteptate. Prin urmare, lucrările viitoare se vor concentra în plus pe evitarea acestor informații demografice, deoarece ar putea avea o părtinire istorică.

Mai mult, este necesară o investigare a cadrului propus în diferite configurații educaționale, de exemplu, folosind acest cadru cu un set de date generat dintr-o clasă fizică.

[2]

Scop:

În lucrarea de față, se urmărește utilizarea a cinci tehnici de învățare automată supravegheată, inclusiv Decision Tree, Naïve Bayes, k-Nearest-Neighbor, Support Vector Machine și Regresia Logistică.

Pentru a analiza efectul unui set de date dezechilibrat, performanța acestor algoritmi este verificată cu și fără diferite metode de reeșantionare, cum ar fi Synthetic Minority Oversampling Technique (SMOTE), Borderline SMOTE, SVM-SMOTE și Adaptive Synthetic (ADASYN).

Metodele Random hold-out și GridSearchCV sunt utilizate ca tehnici de validare a modelului și, respectiv, de ajustare hiper-parametrică.

Este destul de important să se prezică elevii cu performanțe scăzute într-un stadiu incipient, cu o acuratețe mai mare, împreună cu factorii importanți care le pot afecta performanța. Deși există mai multe studii care să prezică performanța academică a studenților, lipsește studiul care ia în considerare toate categoriile de variabile, adică mediul, academic, social și psihologic, și care prezice elevii aflați în risc într-un stadiu incipient, cu acuratețe adecvată. De asemenea, o predicție bazată pe un singur clasificator nu este potrivită dintr-o perspectivă în alta. Mai mult, un clasificator care oferă cea mai mare precizie de predicție pentru un anumit set de date poate să nu fie valid pentru un alt set de date. Astfel, scopul prezentului studiu este de a identifica performanții slabi într-un stadiu incipient, cu o rată de predicție mai mare, prin utilizarea unei abordări scalabile.

Pentru a atinge acest obiectiv, studiul de față are trei obiective importante de cercetare:

1. identificarea caracteristicilor influente prin utilizarea unei tehnici de selecție a caracteristicilor bazate pe filtre.
2. să identifice cel mai performant clasificator prin compararea diferitelor tehnici de învățare automată supravegheată
3. creșterea ratei de predicție a studenților expuși riscului prin utilizarea unui model de ansamblu care integrează cea mai potrivită tehnică de data mining.

Abordare:

Pentru a face datele versatile, acestea sunt colectate de la două colegii de inginerie situate în regiuni diferite (nordul și sudul Indiei). În lucrarea de față, dimensiunea eșantionului cuprinde 550 de studenți de inginerie de la două colegii de inginerie diferite din India, adică Bipin Tripathi Kumaon Institute of Technology, Dwarahat, Uttarakhand, și Cochin University of Science & Technology, Trivandrum, Kerala. Setul de date include informații cu privire la factorii demografici, academici, sociali și psihologici din trecut, cu 27 de atribute diferite. Pentru aceste atribute, datele au fost colectate online cu ajutorul unui chestionar Google Form.

Feature selection:

Selectarea caracteristicilor este o parte importantă a modelului de predicție a performanței studenților din două motive principale:

• Scopul principal al predicției performanței academice a studenților este de a oferi sprijin în timp util studenților cu performanțe scăzute din zona în care acestea lipsesc. Numai după identificarea atributelor care au un impact semnificativ asupra variabilei de ieșire, adică performanța școlară a studenților, pot fi luate măsuri corective adecvate pentru a oferi sprijin studenților cu performanțe scăzute.

• Cu ajutorul selecției caracteristicilor, atributele irelevante pot fi eliminate din date fără a pierde fiabilitatea în clasificare. Astfel, reducerea dimensionalității crește viteza de procesare și, prin urmare, clasificatorul poate învăța mai repede.

În studiul de față, a fost utilizată o tehnică de selecție a caracteristicilor bazată pe filtre, și anume „chi-square”, prin care au fost calculate valorile p pentru fiecare atribut [8]. Atributele cu o valoare p mai mică de 0,01 arată o corelație foarte semnificativă cu notele elevului.

Există diferite tipuri de modele de învățare automată de clasificare care pot fi utilizate pentru a prezice performanța academică a studenților. În studiul de față, au fost aplicate cinci modele unice de învățare automată supravegheată. Pentru a obține cele mai bune performanțe ale acestor modele de învățare automată, parametrii de trecere pentru aceste modele au fost setați cu ajutorul unui algoritm numit „GridSearchCV” care oferă cea mai bună combinație de parametri de trecere [30].

Concluzii:

După aplicarea tehnicii de selecție a caracteristicilor, 11 caracteristici sunt selectate ca fiind caracteristici influente care afectează performanța academică a elevilor. După selectarea celor mai influente atribute, setului de date au fost aplicați algoritmii de învățare automată.

Din lucrarea de față, se poate concluziona că performanțele academice anterioare ale studenților, contextul lor social și demografic și caracteristicile lor psihologice au fost atribute relevante. Astfel, pentru a crește performanța academică a studenților, acești factori pot fi considerați puncte de focalizare.

În studiul de față, toți clasificatorii utilizați au reușit să prezică rezultatele elevilor cu o acuratețe rezonabilă de peste 80%. Dintre toți clasificatorii utilizați, Logistic Regression a fost cel mai performant algoritm, cu un set de date echilibrat și dezechilibrat. În plus, acuratețea și rata de predicție pentru identificarea persoanelor cu performanță scăzută, precum și a celor cu performanță înaltă au fost îmbunătățite atunci când Logistic Regression a fost aplicat la setul de date echilibrat. Precizia predicției a fost îmbunătățită și mai mult cu utilizarea unui clasificator de ansamblu în care trei clasificatoare Logistic Regression au fost integrate cu ajutorul agregării bootstrap. Modelul integrat propus a atins cea mai mare acuratețe de 95,45% și cea mai mare precizie și rechemare pentru performanții slabi cu setul de date echilibrat formulat cu ajutorul tehnicii de reeșantionare SMOTE.

Trebuie remarcat faptul că, cu seturi de date diferite, diferiții clasificatori pot oferi cea mai mare precizie de predicție și, prin urmare, este nevoie ca metodologia să fie scalabilă pentru fiecare situație. Astfel, principalul avantaj al abordării prezente este scalabilitatea sa pentru diferite seturi de date.

Probleme ramase nerezolvate:

Limitarea prezentului studiu este că setul de date examinat are o dimensiune mică a eșantionului și date ușor dezechilibrate, astfel încât în viitor, metodologia propusă ar trebui utilizată cu dimensiuni mari ale eșantionului și date foarte dezechilibrate pentru predicția performanței academice a studenților.

[3]

Scop:

Obiectivul acestei lucrări este de a propune un model predictiv îmbunătățit pentru predicția performanței elevilor.

Selectarea celor mai importante caracteristici este un indicator crucial pentru ca instituțiile academice să facă o intervenție adecvată pentru a ajuta studenții cu performanțe slabe, iar caracteristicile de top care influențează sunt selectate în etapa de selecție a caracteristicilor pe lângă reducerea dimensionalității și construirea unui model predictiv eficient.

Tehnica de clusterizare DB-Scan este aplicată pentru a îmbunătăți performanța modelului predictiv propus în etapa de preprocesare.

Sunt utilizate diferite tehnici de clasificare, cum ar fi Decision Tree, Logistic regression, Naive Bayes, Random Forest, și Multilayer Perceptron.

Prin urmare, obiectivul acestei lucrări este de a propune un model predictiv îmbunătățit pentru predicția precisă a performanței elevilor.

Abordare:

B. Selectarea caracteristicilor

Selectarea caracteristicilor are ca scop selectarea celor mai importante și influente caracteristici din setul de date. De asemenea, este un pas foarte important pentru reducerea dimensiunilor înainte de implementarea metodelor de predicție și clasificare. Am folosit tehnica SelectKBest [15]. Tehnica SelectKBest selectează primele k caracteristici cu cele mai mari valori de scor pe baza testului Chi-Square, pentru a compara rezultatele reale și cele prezise.

Setul de date este colectat dintr-un Learning Management System (LMS) [24]. Conține 480 de înregistrări ale datelor elevilor la diferite niveluri educaționale cu 16 caracteristici.

Caracteristicile au fost clasificate după cum urmează:

• Caracteristici academice: Id-ul secțiunii, Semestrul, Etapele educaționale, Vizualizarea anunțurilor, Nivelurile de clasă, Subiectul, Grupurile de discuții, Resurse vizitate, Mâna ridicată și Zile de absență.

• Caracteristici personale: Sex, Părinte responsabil pentru elev, Părinți care răspund la sondaj și Satisfacția părinților despre școală.

• Caracteristici demografice: naționalitate și locul nașterii.

FROM C EVALUATION MEASURE -> ALMOST EVERYTHING

Concluzii:

Modelul propus este un clasificator de ansamblu de clasificatori Multilayer Perceptron, Decision Tree, și Random Forest. Modelul propus atinge o precizie de 83,16%.

În această lucrare, am propus un model predictiv îmbunătățit pentru performanța elevilor pentru a îmbunătăți acuratețea predicției. Am aplicat diverse tehnici de învățare automată pentru a prezice performanța elevilor. În plus, algoritmul de grupare DB-Scan și pașii de selecție a caracteristicilor au fost exploatați, pentru alegerea caracteristicilor semnificative. Prima noastră metodă de ansamblu a obținut o acuratețe de 83,16%, 78,95% și 65,26% utilizând toate caracteristicile, primele 10 caracteristici de influență și, respectiv, primele 5 caracteristici de influență.

Probleme ramase nerezolvate:

Pentru lucrările viitoare, intenționăm să aplicăm abordarea predictivă propusă la diferite seturi de date, să experimentăm diferite tehnici de selecție a caracteristicilor și să implementăm alternative pentru tehnica de clustering DB-Scan.

[4]

Scop:

Acest studiu estimează succesul elevilor prin analiza datelor din învățământul secundar de la două școli portugheze, ambele incluzând elevi cu niveluri diferite de performanță academică [5].

Scopul principal al acestui studiu este de a ajuta educatorii să identifice elevii care sunt expuși riscului și să-i ajute să îmbunătățească rezultatele educaționale pentru acești indivizi. Sunt folosite mai multe tehnici de preprocesare a datelor pentru a spori acuratețea modelului. Este utilizată o abordare feature subset selection wrapper pentru a localiza cel mai bun subset de caracteristici. Acest studiu explorează, de asemenea, diferențele dintre utilizarea sistemelor de gradare multiple și binare.

Abordare:

2.1 Dataset Description:

Acest studiu folosește două seturi de date disponibile în mod deschis pentru a prezice performanța elevilor. Elevii de la două școli secundare portugheze au produs cele două seturi de date. Pe lângă note, setul de date include informații despre demografia elevilor, statutul socioeconomic și școala. A fost dobândit prin rapoarte școlare și sondaje. Datele de la o lecție de limba portugheză sunt incluse în al doilea set de date, inclusiv o clasă de matematică. Există 33 de variabile în total în ambele seturi de date.

S-a decis că vor fi folosiți trei algoritmi de clasificare bine-cunoscuți pentru a prognoza notele elevilor. Acești algoritmi au fost DT, RF, NB, MLP și JRip [7].

O operație de preprocesare clasifică atributul de nota finală în două sisteme de notare independente. Acum există două versiuni diferite ale ambelor seturi de date. Numeroase formate de notare (inclusiv cele binare) au fost disponibile atât pentru seturile de date matematice, cât și pentru cele portugheze. Din această cauză, putem compara rezultatele numeroaselor posibilități. Folosind setul de date portughez în versiunile sale multiple de gradare și binare, primul experiment a comparat cinci algoritmi. Tabelul 3 arată cea mai mare rată de acuratețe de 75,40% pentru versiunea de gradare multiplă a acestui set de date [8].

O rată de precizie îmbunătățită a fost obținută folosind metoda Random Forest cu gradarea binară. Potrivit unui set de date, în care notele finale sunt clasificate ca Reușit/Eșec, procentul de acuratețe a crescut la 95,08%. Algoritmul Decision Tree a oferit cele mai bune rezultate pentru sistemele de notare multiple, cu o rată de precizie de 85,62%. La 93,49%, cea mai excelentă rată de acuratețe a versiunii setului de date binare a fost atinsă de Random Forest.

După preprocesarea setului de date sau alegerea unui subset de caracteristici într-un al doilea experiment, am efectuat toate comparațiile. Metoda subsetului de wrapper a selectat cele mai relevante atribute și a crescut ratele de precizie. Este esențial să alegeți caracteristicile potrivite pentru ca un model să reușească. Selectarea calităților adecvate și eliminarea celor redundante este un proces simplu în doi pași. În primul rând, selecția atributelor este făcută pentru a dezvolta un model de bază, un model mai ușor de citit și pentru a decide care atribute sunt cele mai relevante pentru constatări. Apoi, am folosit filtre și wrappers pentru a alege caracteristicile. Metoda wrapper a fost folosită în această investigație deoarece produce în mod constant rezultate mai bune. Această procedură folosește o structură recursivă. Invocarea algoritmului pe un subset este primul pas al procesului. Evaluarea se bazează pe succesul modelului. Această examinare are două rezultate posibile. Dacă alegeți opțiunea unu, va trebui să începeți din nou de la zero; șansa doi vă permite să utilizați ceea ce ați ales deja [9]. Rata de precizie a unui algoritm de arbore de decizie a fost crescută de la 70,80% la 80,56% prin utilizarea abordării wrapper subset. Rata de precizie a algoritmului Random Forest a crescut de la 82,40% la 85,74%. A existat o creștere a preciziei de la 77,79% la 84,56% folosind abordarea Naïve Bayes. Rata de acuratețe a abordării Multistrat Perceptron a crescut de la 73,60% la 78,45%. Procentul de acuratețe al metodei JRip a crescut de la 75,30% la 80,87%.

Concluzii:

Conform constatărilor acestui studiu, este posibil să se prognozeze notele finale ale studenților folosind abordări de extragere a datelor. Trei algoritmi de clasificare bine-cunoscuți, cum ar fi Decision Tree, Random Forest, Nave Bayes, Multilayer Perceptron și JRip, au fost testați pentru ratele de precizie. Strategia de selecție a subsetului de caracteristici wrapper a fost utilizată pentru a îmbunătăți performanța clasificării. Metodologia de selecție a atributelor wrapper-ului a crescut semnificativ acuratețea în toate metodele.

Acest raport sugerează că notele finale ale studenților pot fi prezise folosind tehnici de extragere a datelor bazate pe cercetări anterioare. Pe două seturi de date educaționale legate de orele de matematică și lecțiile de limba portugheză, trei abordări binecunoscute de extragere a datelor, cum ar fi Decision Tree, JRip, Naive Bayes, Multilayer Perceptron și Random Forest, au fost utilizate în experimente. Ca rezultat, folosind metodele de extragere a datelor utilizate, succesul studenților ar putea fi prezis cu o acuratețe rezonabilă.

Probleme ramase nerezolvate:

În viitor, pot fi folosite diferite metode de selecție a caracteristicilor. În plus, seturile de date pot fi supuse unei varietăți de metode de clasificare.

[5]

Scop:

În acest studiu, am examinat comportamentele elevilor în sarcina de autoevaluare online și modul în care le afectează performanța de învățare.

Abordare:

Un experiment de 6 săptămâni a fost efectuat într-un curs de contabilitate. Studenții au fost instruiți să finalizeze autoevaluarea online sub formă de chestionare formative după oră. Am efectuat o analiză de grupare, care a evidențiat trei modele de comportament în autoevaluarea online și am comparat performanța de învățare a elevilor în diferite modele.

3.1. Participants and context

A 6-week experiment was conducted for students studying in an accounting course at the Accounting department of a Taiwanese university. The course is mandatory for fourth-year students in the Ac counting department, and they all have the same educational background and the same exposure to information technologies. The experiment was conducted from the beginning of the semester to the midterm examination. The course employed BookRoll, an e-book reading tool developed by Kyoto university, where the instructor uploads learning materials before each class and students can perform various activities during their reading. For an introductionto the e-book reading activities available on BookRoll, the studies of Ogata et al. (2015) and Flanagan and Ogata (2018) can be referred. In this course, the Moodle online learning environment was used together with face-to-face teaching. The activities in Moodle included accessing learning materials, discussing with peers or a teaching assistant, and taking assessments in the form of formative quizzes. Most students have used Moodle before, but none of them have experience with BookRoll and the assessment system employed in this study. Therefore, students were introduced to Moodle, BookRoll, and the assessmentsystem as well as how to use them at the beginning of the experiment. The assessments were generated every week and consisted of cloze items related to the content taught during that week. Since the assessment was designed to assess students’ ability to recall the concepts in the study materials, it mainly contained factual questions. The questions were generated by the online assessment system proposed in a previous study (Yang et al., 2021). The system extracts keywords from the text of relevant learning material and masks those words to form the cloze items. To ensure that the masked items reflect students’ knowledge level on the learned.

subjects, the questions were reviewed by the instructors before being presented to students. The instructor could modify the cloze items by adding additional masks or removing existing masks to ensure that the items fit the course objectives. Students were asked to take the assessments after classes. Fig. 1 shows a snapshot of using the online assessment system. After entering the system, students can click the green mask to type and submit their answer. The system will provide immediate feedback by showing the result of their answer. The mask is removed if the student provides the correct answer. Otherwise, a red “x” is displayed in the input field. Students can choose to see the hint by clicking the “Hint” button if they cannot recall the answer. To discourage students from finding answers from learning materials during the assessments for achieving high scores, they were informed that their performance in the assessments would not affect their final grade. Thus, students repeatedly took the assessments for review purposes. During the experiment, five assessments with distinct properties were generated. The number of questions in each assessment was 15, 11, 11, 13, and 8, respectively.Students could take the assessments through the Moodle anytime and anywhere. Student self-assessment behaviors were recorded in the database with timestamps. The actions that students can perform in the assessments are presented in Table 1. The relationship between students’ self-assessment behaviors and their learning performance was analyzed in this study. The midterm examination scores for the course were used as the indicator of learning performance. The midterm examination involved the topics covered in the five assess ments, but the contentwas presented through different questions. With a total score of 100, each topic consisted of 20%. A total of 73 students participated in this experiment. Seven students did not take the assessment at all and werethus excluded. All students were informed that their self-assessment behaviors and learning performance would be logged, and that their personal information would be anonymized for research analysis. All participants agreed to the policy. 3.2. Feature selection and data analysis A total of 88696 online self-assessment events were collected and analyzed. The features presented in Table 2 were extracted from the logs. The attempt count was used to measure how frequent a student took the assessment. Question views indicated the number of times a student clicked and viewed the questions during the assessment. A high number of question views indicated that the questions were likely reviewed multiple times. The submission rate measured how many attempts students required before obtaining the correct answer; that is, how easily they were able to recall the learning content. The hint rate was used to evaluate students’ reliance on hints. The average values of attempt count, question views, submission rate, and hint rate were6.79, 14.23, 2.47, and 1.76, respectively. We applied hierarchical clustering to identify specific patterns of online self-assessment behavior for answering RQ1. The number of cluster was determined with the help of dendrogram that shows the distance between each data point. Since each feature has a different scale, which may affect the performance of clustering algorithms, all features were standardized for clustering analysis. Next, the difference between the learning performance of students from the identified patterns were examined using statistical model for answering RQ2. Data analysis and data visualization were performed using Python. Specifically, clustering analysis was performed using hierarchical clustering through the Scikit-learn package. Statistical analysis was carried out using the SciPy package.

Because the distribution of each feature did not satisfy the requirement for performing ANOVA, we employed the Kruskal–Wallis H-test to examine whether the performance in each feature differed significantly between clusters. The number of features in each cluster is presented in Table 3. The resultsindicated a significant difference in the frequencyof using the test module (H = 25.18, p < .001), question views (H = 28.30, p < .001), submission rate (H = 34.26, p < .001), and hint rate (H = 18.78, p < .001) among the three clusters. Students in Cluster C (median [M] = 9, standard deviation [SD] = 2.47) and Cluster B (M = 7, SD = 2.31) used the assessment system significantly more often than did those in Cluster A (M = 5, SD = 2.28), and students in Cluster C also had more assessment attempts than did students in Cluster B. The Mann–Whitney post hoc test results indicated that Cluster B students (M = 21.5, SD = 5.54) viewed significantly more questions than did Cluster A (M = 9.7, SD = 5.09) and Cluster C (M = 13.9, SD = 3.95) students, and Cluster C students also viewed more questions than did Cluster A students. Moreover, Cluster B students (M = 4.31, SD = 1.27) exhibited a significantly higher submission rate than did Cluster A (M = 1.42, SD = 0.94) and Cluster C (M = 1.68, SD = 0.65) students. Finally, the post hoc test results indicated that Cluster B students (M = 3.03, SD = 1.46) viewed hints significantly more frequently than did Cluster A (M = 1.69, SD = 1.73) and Cluster C (M = 0.1, SD = 0.78) students, and Cluster A students also used hints more often than did Cluster C students. The findings showed significant differences in behavior within each cluster, suggesting that it is feasible to use hierarchical clustering to classify students’ self-assessed behavior. To understand how individual features affect learning performance, we performed Spearman correlation analysis to analyze the relationship between each behavior and the examination score. The attempt count was positively correlated with learning performance (r = 0.3, p < .05), whereas the feature of question views was not (r = 0.15, p > .05). This indicates that assessment taking attempts in itself can generate positive effects for learners, and the frequency of viewing questions is not as critical as assessment taking attempts. On the other hand, submission rate (r = 0.24, p < .05) and hint rate (r = 0.29, p < .05) had a negative impact on learning performance, indicating that high submissions and the frequent use of the hint function may hamper student learning.

Concluzii:

Concluziile studiului actual oferă perspective cercetătorilor din domenii conexe și practicienilor din domeniul educației. Cercetătorii pot optimiza instrumente similare de evaluare online pe baza rezultatelor noastre; pot colecta mai multe caracteristici care reflectă comportamentul elevilor și pot vizualiza rezultatele analizei comportamentale pe un dashboard pentru ca instructorii să monitorizeze procesul de învățare al elevilor. Rezultatele noastre pot servi drept referință pentru instructorii care doresc să înțeleagă comportamentele elevilor și dacă există modele specifice în timpul autoevaluărilor. De asemenea, instructorii pot folosi rezultatele noastre pentru a identifica dacă elevii demonstrează comportamente nestandard în timpul autoevaluării.

Rezultatele au indicat că studenții care au susținut frecvent evaluările online după ore au avut tendința de a obține un scor la examen mai mare decât cei care nu au făcut-o. Cu toate acestea, performanța de învățare a elevilor care au demonstrat comportamente nestandard nu s-a îmbunătățit neapărat, deși au luat parte activ la evaluări. Rezultatele noastre sugerează că comportamentul elevilor este un factor critic în îmbunătățirea învățării prin autoevaluare. Aceste constatări oferă informații pentru cercetătorii din domeniul analizei învățării, precum și pentru practicienii care doresc să adopte autoevaluarea online pentru învățare.

Probleme ramase nerezolvate:

Studiul nostru are mai multe limitări.

Dimensiunea mică a eșantionului (76 de studenți) a redus generalizarea rezultatelor noastre. Sunt necesare rezultate similare din experimente pe scară largă pentru a confirma relația dintre comportamentele de autoevaluare online ale elevilor și performanța de învățare.

Adăugarea mai multor funcții ne poate permite, de asemenea, să obținem informații mai detaliate și mai precise asupra comportamentului elevilor în timpul autoevaluărilor online. De exemplu, înregistrarea timpului de răspuns al fiecărui element ne poate ajuta să identificăm comportamentul de ghicire.

Mai mult, studiul actual a analizat comportamentele de autoevaluare ale elevilor folosind o abordare bazată pe frecvență, care poate neglija relația detaliată dintre caracteristici. De exemplu, dacă elevii văd indicii înainte de a răspunde poate avea ca rezultat două modele. Prin urmare, o altă abordare analitică a învățării, cum ar fi procesele miniere, ar trebui luată în considerare pentru a explora alte modele comportamentale de autoevaluare online în studiile viitoare.

În viitor, vom efectua o analiză comportamentală mai detaliată a comportamentelor de autoevaluare online, utilizând alte abordări, cum ar fi analiza temporală, care a fost aplicată în alte studii (Kokoç et al., 2021), pentru a investiga dacă comportamentele elevilor se schimbă în timp. Analiza temporală este o altă abordare care poate determina dacă elevii urmează același model în timp ce fac evaluările pe parcursul experimentului.

În cele din urmă, compararea abilităților SRL, a strategiilor de testare și a diferențelor individuale cognitive ale elevilor care prezintă modele comportamentale similare poate dezvălui informații valoroase. Studiile viitoare care modelează comportamentele de autoevaluare online ar trebui să țină cont de acești indicatori.

Cercetătorii pot aplica clustering analysis pentru a examina comportamentul elevilor în alte sisteme de autoevaluare online care oferă diferite caracteristici pentru a identifica mai multe modele comportamentale. În plus, profesorii pot oferi feedback personalizat elevului individual, pe baza modelelor comportamentale ale acestora, pentru a-i ajuta să maximizeze beneficiul autoevaluării.

[6]

Scop:

The objective of this article is to show the implementation results of a predictive information system (IS) for the prevention of university dropout in a higher education institution. The system allows the calculation of the risk of dropout per student and uses an alert generation procedure to coordinate interventions.

Abordare:

This article aims to show the results obtained through the implementation of a predictive information system (IS) for the desertion of university students. The methodological design consisted of 3 phases: data collection, modeling, and implementation/validation. The resulting IS was implemented in a higher education institution with more than 15,000 students, provides a user interface for monitoring the main risk factors associated with student’s dropout, in a timely and personalized way.

The data was collected, consolidated, and complemented with information is obtained from the Colombian Ministry of National Education, and the Colombian Directorate of Social Development). 2.2. Phase 2: Modeling. 2.2.1. Variable Selection. Given that the students of each higher education institution have a specific set of characteristics, the dropout patterns may vary according to the institution. A “naive” approach was adopted, aiming to achieve an objective point of view at the time of analyzing the dropout phenomenon. All data were included in the analysis, without any of the variables being previously discarded. This selection process aims to reduce the number of variables that will be used to train the model, prioritizing those with the most significant predictive potential. 2.2.2. Classifier Selection: Instead of predefining a single model that works for most educational institutions, several methods were tested simultaneously, and finally, the method that best suits the reality of the institution was chosen. Thus, the predictive model will be specifically adapted to the dropout patterns observed in the students of the institution. The goal of this step was to determine the classification algorithm with the highest prediction efficiency on the dataset. The algorithms considered during this phase included: AdaBoost algorithms, Bayesian GLM, Decision Trees, Logit Boost algorithms, Random Forest, and Stochastic Gradient Boosting.

3. Results 3.1 Variable Selection The data collected in Phase 1 consisted of 44,031 records of 15,805 students, covering four academic periods (2016-1, 2016-2, 2017-1, and 2017-2), grouped in 165 different initial variables. A correlation analysis was carried out using the Kolmogorov Smirnoff (KS) test for each of the variables related to the dropout rates. Subsequently, a multivariate analysis method was used using a Random Forest Algorithm to determine the degree of contribution of various sets of variables in the prediction of dropout rate. Decision trees were generated for the most significant variables. The decision trees provided the ranges or values where the dropout rate is significantly lower or higher than the average (Figure 1). The students were classified in one of the branches (nodes), and in the row depending on the category in which the student is (1: drop out; 0: remain).

All the variables were ordered according to their predictive potential. This process requires the use of advanced Data Mining techniques because, for each variable, its conditional dependence on the presence of other variables with respect to dropout must be taken into consideration, and not only its non-conditional dependence. A Random Forest algorithm was used as a regression method on the data (not to be confused with Random Forest in its use as a classification algorithm, which is explained in the classifier selection stage). The process consisted of the following stages:

• Random groups are chosen from the variables present in the dataset, where each variable can be included in more than one tree, and the total of variables per group can vary.

• For each group, a random subset of rows is chosen from the original dataset (student enrollment).

• A decision tree is trained for each group, predicting dropout with the different training data.

• Each variable receives an importance value, which is measured considering the set of decision trees with and without it.

• The variables are ordered according to their importance, and a multivariate ranking is obtained.

3.2. Classifier Selection.

The procedure for the classifier selection started choosing the first variable in the multivariate ranking and training a model for each type of algorithm, predicting dropout based on that single variable. The precision result was measured according to the area under the (Receiver Operating Characteristic Curve) ROC, and the result of each partial model was saved. Then the procedure was repeated with the first two variables in the ranking, then the first three, and so on. The first 30 variables were used at most, to avoid overfitting and to reduce the number of variables that will enter the final result. The different classifiers tested in this stage are described below:

• AdaBoost Algorithm:

AdaBoost is an algorithm to build a “strong” classifier from the weighted sum of a large number of base classifiers, and by default, they are decision trees. AdaBoost classifiers generally have lower predictive quality, but they are easier to obtain. In each iteration, the algorithm learns from the misclassified data, updating the weights that accompany the weak classifiers, and thus improving the predictive quality at each stage [13,14].

• Bayesian GLM: In the same way as the Logistic Regression, this model focuses on predicting the occurrence of a binary (or dichotomous) phenomenon. However, the parameters of this model are random variables so that probability distributions can be assumed for them before the arrival of data, and the change in their distribution caused by the data considered in the analysis can be known [15].

• Decision Trees: A tree is made up of a set of nodes, branches, and leaves. Each node represents a variable, which can be divided into branches or leaves. The main decision rule is to separate the data into two groups, as different as possible and to do this, it looks for the variable that best performs this segmentation. All of this process continues until a constraint is met: it can be a fixed number of nodes or fixed depth of the tree. To predict using this method, a “new” student is taken, and how he moves inside the tree is observed, reaching a final decision, which in this case is whether to drop out or not [16].

• LogitBoost Algorithm: This algorithm uses the same principle as AdaBoost, but the base classifiers used in this process are classifiers created from logistic regressions [17].

• Random Forest Random Forest algorithm creates a large classifier formed from the development of many small decision trees created randomly. The final prediction is made through the weighted average of the classification of all trees. [18].

• Stochastic Gradient Boosting: This technique consists of iteratively creating a classifier that, using regressions obtained from sub-samples of the data, learns at each stage of the data that is misclassified, adapting to the observed reality [19].

The area under the ROC curve provides a measure of the performance of the model for the selected variables, equivalent to the probability that the model classifies a real dropout student with a higher probability than a false dropout student (false positive). The two algorithms with the largest area under the ROC curve (close to 80%) was the AdaBoost algorithm and Bayesian GML and provided the best balance between predictive level and model stability. From the initial 165 variables, the first 30 most significant were used to select the classifier. Finally, a set of 25 variables allowed the model to obtain a precise prediction. The probability of dropping out of each student was calculated using the 25 variables. This probability defines the level of priority that will be taken when carrying out activities to monitoring and preventing dropouts by those responsible for the follow-up process. The probability changes according to the alerts and interventions registered by the responsible for the follow-up process during the academic period (Section 3.3.2. and 3.3.3.). At the end of each semester, the impact of the interventions on the percentage of student retention is calculated.

Concluzii:

The implemented IS represents a powerful tool for predicting, monitoring, and managing the risk factors associated with the student dropout factors. Some advantages of the implemented IS include the centralization of the information allowing a comprehensive view of the students, the prioritization in the timely follow-up of students with

a higher risk of dropping out. Also, it allows individual and massive register of the interventions carried out to students. This includes scheduling of future interventions and defining a way to evaluate of the impact of follow-up strategies on student permanence.

Probleme ramase nerezolvate:

The limitations of the IS include the need to make more flexible the assignment of a higher number of monitoring managers to generate alerts and improving the cu stomization of reports. An update of the software is proposed that incorporates new functionalities in terms of bulk uploads, increasing data volumes when downloading reports, and improvements in the implementation of the strategy by those responsible, giving priority of intervention to students according to the level of risk using the following order 4, 3, 5, 2 and 1 for greater effectiveness. Among the opportunities for future research is the incorporation of new classification and weighting methods, which allow to improve the reliability of the information system predictions.

[7]

Scop:

The study aimed to characterize learners according to their learning patterns and to identify indicators that predict students’ success in an online environment.

Abordare:

Specifically, we focused on the role of a central factor affecting success in online courses: self-regulated learning and learner engagement. To this end, we used a mixed methods approach that combines semi-structured interviews and statistical analysis. We applied two logistic regression models and a decision tree algorithm and found two parameters that can predict completion of the course: the submission status of an optional assignment and the students’ cumulative video opening pattern (SCOP). Recommendations for institutions and lecturers regarding the benefits of implementing these models to identify self-regulated learning patterns in online courses and to design future effective interventions are discussed. Regarding students, we emphasize the importance of time management and how choices they make with respect to their learning process affect their potential for success.

Here we present a study on three online undergraduate general chemistry courses offered at the Open University of Israel (OUI). Existing research on predicting persistence in chemistry courses focused on background indicators, such as high-school achievement and SAT scores (Lewis & Lewis, 2007). However, the rise of online platforms (Amaral, Shank, Shibley Jr, & Shibley, 2013) opened an opportunity to focus on more proximate indicators that can predict student performance in each course and extend the analysis, beyond the students’ past achievements and background, by analyzing their actual online learning patterns.

As discussed above, completing online courses is known to be more difficult than traditional face-to-face courses. Our study has two main goals: (1) to characterize learners according to their learning patterns in the online learning environment and (2) to identify learning patterns that can predict students’ successful completion of online chemistry courses. Because online course data generally present information about learning behavior, this study includes measures that relate to the frequency of online lesson openings and assignment submissions, which are indicative of the extent of learners’ engagement and SRL. To meet this goal, we posed two research questions:

1.What learning behaviors do students apply in online general chemistry courses?

2. What indicators of online learning can predict course success, and at what stage?

We studied undergraduate general chemistry courses offered at the OUI, which does not have prerequisite admission requirements for undergraduate degrees. Nevertheless, the admitted students need to demonstrate high levels of knowledge and skills in order to successfully complete the courses. To maintain this kind of accessibility, the OUI offers a variety of learning tracks that allow students to choose between traditional face-to-face classroom, online learning, or a combination of the two (blended learning). Such flexibility provides more opportunities for potential students from the country’s geographic and socioeconomic periphery, as well as other students that can benefit from flexible academic schedule (https://www.openu.ac.il). Here we followed students who took the chemistry courses in an online format.

The research design included both qualitative and quantitative tools (mixed methods). Integrating quantitative and qualitative research methods is known to increase the precision and trustworthiness of the results (Leech & Onwuegbuzie, 2007). The study was conducted in two stages: The first stage was descriptive and the second was predictive. The purpose of the first stage was to identify learning behaviors that students apply in the online general chemistry courses. This stage was based on interviews with participants enrolled in each of the these courses, and a descriptive analysis of the courses’ log files. Before deciding which variables to use in the quantitative analysis, we conducted a preprocessing phase of data collection, categorization, and filtering (see Section 4.2). Then, we found that the reliable features that can be used are the activity time stamp, video play, demographic data, and the assignment submission status. In addition, we considered the course pedagogy, which enabled students to choose which assignment to submit (see Table 1), and the flexibility of the online course, which allowed students to play the video at their own time and pace. Following this stage, the predictive stage began, during which we quantitatively evaluated the collective power of the major variables that were identified in the first stage as predictors of course success.

4.1.Qualitative analysis In order to reveal students’ learning habits, we conducted 13 semistructured interviews with participants enrolled in each of the three courses included in this research. Semi-structured interviews aim to enable flexibility. The interviewer prepares a list of topics and questions and follows that list during the interview. In order to ensure that the questions elicit open responses, the interviewer enables the responses to be developed in ways that were not anticipated when the interview was planned (Brown & Danaher, 2019). Out of the 13 interviews, 10 were conducted with students who successfully completed one of the three courses analyzed in this study, whereas the other three were with students who did not complete the course they took. Twelve interviews were conducted by phone, and one in a face-to-face meeting. In each semester we posted an advertisement in in one course website (alternating between the three courses throughout the year) and invited volunteers to be interviewed at the end of the course, after the final exam. The semi-structured interview protocol included twenty two questions organized around subthemes (see the Supplementary Materials file). All three authors validated the interview’s questions. Each interview lasted 20–60 min and was audio recorded and transcribed.

4.2.Quantitative analysis The quantitative analysis covered data from Moodle log files, course grades, and students’ demographic profiles. The Moodle log files contained the course activity reports, which showed the number of views for each course website resource. The grades (including the assignment submission data) and demographic data included a complete set of student characteristics (from a particular semester), such as the district of residence according to the socio-economic background, gender, educational background, achievements, and the assignment submission status. Each Moodle log file contained data about a course in a particular semester and included a free text column that described an action performed by a Moodle user, identified by a Moodle ID string. In order to follow research ethics principles and to safeguard students’ privacy and in accordance with the EU’s General Data Protection Regulation (GDPR) and Israel’s Protection of Privacy Law, prior to transferring any data to the researcher, identifying fields such as the given name and surname were removed. In addition, national ID numbers and Moodle identifiers were encrypted. The research received IRB (Institutional Review Board) approval by the OUI Ethics Committee.

4.2.1. Quantitative collective variables The database described above listed numerous instantaneous activities of individual students in specific courses. In order to analyze these data in a meaningful way, we defined two collective variables, listed below, which describe students’ learning patterns. Further details regarding the rationale for developing these variables are presented in the Results section. a) Student Cumulative Opening Pattern (SCOP): The SCOP represents cumulative video sessions that each student opened until a given week, disregarding a repeated opening by the same student (discrete interval variable). For example, if student A opened one video in the first week, another one in the second week, and none in the third week, then by the third week his SCOP would equal 2. We used the SCOP to estimate learners’ progress in the course throughout the semester. b) The submission status of the first optional assignment: This is a binary variable that distinguishes between students who submitted the first optional assignment and those who did not. A Shapiro-Wilk test of normality distribution was statistically significant, indicating a univariate normality deviation (Villasenor Alva & Estrada, 2009). Therefore, we used chi-square and Mann-Whitney U tests to examine the association between the measured parameters and students’ success rate (MacFarland & Yates, 2016; Onchiri, 2013). 4.2.2. Logistic regression In order to address the research questions and predict whether a student is going to succeed in a course, there is a need for a statistical methodology that could explain a dichotomous outcome (successful/ unsuccessful) based on a collection of categorial, ordinal, and interval independent variables. The logistic regression approach (Osborne, 2015) provides such an analysis by moving from predicting an event occurrence to predicting its probability to occur. This type of analysis was used successfully in a number of educational studies (Artino Jr & Stephens, 2009; Yair, Rotem, & Shustak, 2020). In our research, we followed the user-controlled method of entry approach described by Osborne (2015), in which the researcher decides which variables to include, and they are all entered simultaneously. According to Osborne (2015), this is theory driven, meaning that the analysis is based on prior theory or research and is therefore more defensible. The method involves defining a new dependent variable, Logit: Logit = Ln(p/(1 p)) , (1)

where p is the probability for an event occurrence, here – success in the course (Osborne, 2015). The Logit function is estimated by a regression model, which is a linear function of a set of independent variables {Xk} with coefficients {bk}: Logit(Y)= b0 + b1X1 + ... + bkXk (2) Here Y is the dependent variable, b0 is the intercept, and {bk} measures the slopes, or the effects with respect to {Xk}. Logistic regression is a commonly used prediction algorithm with strong predictive performance and good comprehensibility. Despite these strengths, logistic regression results can be biased due to interaction effects between variables or multicollinearity. To address this weakness, we also used a decision tree algorithm (De Caigny, Coussement, & De Bock, 2018).

4.2.3. Decision tree algorithm

A decision tree (DT) is a data mining technology that carries out various classification and regression tasks and is generally considered effective and simple (Lin & Fan, 2019). Starting from a parent node, DTs involve a recursive process of splitting nodes into smaller and purer subsets by iterative determination of optimal splitting criteria that divide the data over two child nodes. This process terminates when no further splits are desirable or possible (De Caigny et al., 2018). Here we used the chi-square Automatic Interaction Detection (CHAID) (Kass, 1980) method as an attribute selection measure, based on the statistical chi-square test for independence. Following Hershkovitz and Nachmias (2011), we applied a significance level of 0.05 for splitting nodes and merging categories, and 10-fold cross-validation. CHAID is considered to have a high defection prediction accuracy (Lin & Fan, 2019). It is mainly used to calculate the degree of dependence between several variables – the larger the value calculated by chisquare, the higher the degree of dependence and the probability value of the variable. Moreover, a probability value is used to determine whether to continue the splitting process in the CHAID algorithm to estimate all the possible predictive variables. The significance levels of the differences between the various categories of dependent variables are tested for each variable. Then, the insignificant categories are merged into a homogeneous group, and the remaining categories are analyzed repeatedly until the differences are no longer significant.

Let us start with a description of the interviews, which represents the first qualitative analysis stage. We used the interviews to characterize students’ learning behavior in the online chemistry courses according to the SRL framework. They also helped us identify the main variables for the regression model. Out of the 13 interviewees, 10 successfully completed one of the three courses analyzed in this study, whereas the other three did not complete the course they took. Fig. 1 presents a summary of the interview analysis in the form of a heatmap of the 33 mapped categories that represent students’ SRL characteristics (see Methodology, Section 4). These categories are divided according to the six main SRL dimensions of Barnard et al. (2009): (1) goal setting, (2) environment structuring, (3) task strategies, (4) time management, (5) help-seeking, and (6) self-evaluation. The heatmap’s scale shows the frequency that each category appeared in our interviews. To prevent bias, we counted each category once for each interview, even when it emerged multiple times, such that high frequency refers to categories that emerged in different interviews. This analysis allowed us to examine SRL in the context of online education in chemistry, since students often described their difficulty with chemistry content.

6.2.1. Assignments submission

At the OUI, students receive the assignments and their submission schedule in advance before the semester begins. The teaching staff of all three courses made the first assignment mandatory, to take advantage of the students’ motivation at the beginning of the course and to create a commitment for learning. Hence, it is not surprising that most of the students submitted it, and therefore, it could not be used as a predictorof course success. Likewise, the other 2–3 mandatory assignments in each course had a relatively high submission rate. In addition, each course included 2–3 optional assignments, of which the students had to submit at least one. Naturally, those who submitted more optional assignments had a higher probability to succeed in the course (Fig. 2), thus making the submission rate of optional assignments a good parameter that distinguishes between those students who successfully completed the course and those who did not. Of these, the first optional assignment (which was the second assignment in all three courses, with a submission deadline at the 5th week) had a much lower submission rate, compared with the first mandatory assignment, which made it an informative variable for predicting success at an early stage of the course. Table 5 shows the percentage of students that submitted the first two assignments. 6.2.2. Video sessions’ opening pattern The courses in this study consisted of 12 online sessions, which students can view either live (synchronous) or recorded (asynchronous). Since many students did not participate in the live sessions and opened the recorded sessions asynchronously, we did not distinguish between synchronous and asynchronous video opening. Note that, as with most online generated data, we know whether a student clicked and opened a video, but we have no way of knowing whether the student actually viewed the entire session. Therefore, we referred to this as an opening pattern and not as a viewing pattern. Fig. 3 shows unified data from all the courses; it counts the number of students who opened each session throughout the semester. The colors indicate two groups of students: (1) those who succeeded in the courses and (2) those who did not complete them. Here we counted each student once per session. As can be seen, the first group shows a steady pattern of sessions that opened – the number of students is constant throughout the semester, and almost all of them opened each video session at least once. On the other hand, the second group of students did not follow a steady pattern, and the number of students who opened each session significantly decreased throughout the semester. Fig. 4 presents a different view of these data; the percentages of students from each group that opened the sessions’ first, second, third, and fourth quartiles are counted. As is evident, almost all the students in the first group, who successfully completed the courses, played the entire set of tutoring sessions. Most of the students who did not complete the course opened only some of the sessions. Based on these results alone, we could not determine whether students who did not complete the course decided to use other course learning materials. Nevertheless, all learning materials were available to all the students, to begin with. Figs. 3 and 4 provide information about the opening patterns that accumulated throughout an entire semester. Both figures help distinguish between students who successfully completed the course and those that did not. A close look at Fig. 3 shows that the number of students who opened the first few video sessions was approximately similar between the two groups. Our original goal was to detect the course completion status during the semester. Therefore, we needed a more accurate parameter that could distinguish between the two groups at the early stages. From the interviews, we learned that there were students who successfully completed the course and watched the online sessions from week to week. The SCOP variable (see Section 4.2.1), which counts the total number of different video sessions that each student opened up to a specific week, was used to analyze different learning patterns. We designed the SCOP variable to disregard multiple plays of the same video. In this way, it represents students’ progress with the course material throughout the semester. For example, if learners’ SCOP equals 9, we know that they played nine of the course’s videos. If we had counted multiple video plays, we would not know whether the number 9 in dicates that the student progressed in the course material or stopped to repeat specific lessons. Fig. 5 presents the weekly average SCOP for each group (who successfully completed and did not complete the course). As is evident, this parameter is quite informative for distinguishing between the two groups, even at earlier stages of the course. Note that the course lasted for 14 weeks. Data for weeks 15–20 represent the exam period. It is included here to show that students continued to open the video sessions at high rates toward the exam date. The group of successful students used the video resources much more than the other group did. In order to assess the statistical association between the SCOP and the course success, we conducted a Mann-Whitney U test of the SCOP distributions among students who completed the course and those who did not. We found that starting at the second week and throughout the rest of the semester, the first group displayed statistically significantly higher scores in the Mann-Whitney U test than the second group (p < 0.05), teaching that this variable is a good candidate for predicting success rates. 6.2.3. Logistic regression models Based on the results presented above, we defined two main independent variables for the logistic regression model: the first optional assignment submission status and the SCOP. Prior to the logistic regression analysis, we conducted a correlation analysis for the set of independent variables in order to test for multicollinearity. Two categories of independent variables were used as control variables in the analysis. The first are demographic variables that included gender and the district of residence. The age variable was not used, since we found multicollinearity between this variable and both the advanced diploma and the district of residence. The second category is educational background, which included prior and current studies: the existence of a prior advanced diploma, an indication whether the current course is the first course at the OUI, the semester index, and the course name. These variables are suitable as controls, since although they may influence students’ success in the course, they do not change during the semester and do not depend on students’ learning choices during the course. No significant associations were found between any of the control variables, as well as the variables of assignment submission status and the SCOP. However, we did find multicollinearity between the assignments’ submission status and the SCOP (namely, these variables are correlated). Therefore, we ran two different models of logistic regressions for each of them. Then, we carried out a logistic regression analysis using SPSS version 24 (IBM Corp., 2016) and R version 4.1.2. For each model we chose which variables will be included in the model and they were all entered simultaneously. 6.2.3.1. Logistic regression analysis of the course achievements. Logistic regression models were built based on the data collected from all three courses. Each model was based on the data of students enrolled in the years 2016–2019 (n = 797). Data from the first semester of 2020 (n = 157) were used to validate these models. The course name was used as a control parameter. Model A used the submission status of the first optional assignment, which was the second assignment in all courses. Among the 797 students, 478 (60%) submitted the first optional assignment. A chi-Square test found a statistically significant association between the first optional assignment submission status and the overall course success (χ(1) = 129.49, p < 0.000). The effective size of this finding, Cramer’s V, was moderate (Kotrlik, Williams, & Khata, 2011) and significant (φ = 0.403, p < 0.000.). As shown in Table 6, most of the students who successfully completed the course submitted the first optional assignment, whereas most of the students who did not complete the course did not submit it. These results justified building the model based on the first optional assignment submission rate. The results of model A are presented in Table 7. The Wald statistic, defined as the square of a regression coefficient divided by the standard error of that coefficient (Osborne, 2015), was applied to determine the statistical significance of each independent variable. The logistic regression model for the entire sample (n = 797) was found to be statistically significant χ2(6) = 129.079, p < 0.001. Following our expectations described above, the submission rate of the first optional assignment (p < 0.01) was found to be a significant parameter for predicting the final course success status, along with the advanced diploma (p < 0.05). The course name was found to be insignificant, justifying the analysis of the three courses as a single database. The model correctly classifies 70% of the cases (see the model evaluation). Furthermore, using the Hosmer-Lemeshow Goodness-of-Fit tests (Hosmer & Lemesbow, 1980; Paul, Pennell and Lemeshow, 2013), we found that our regressions are assessed to be well fitted. These results suggest that when students submit their first optional assignment, we can determine the probability that a specific student will complete the course. Model B’s results, which are based on the SCOP variable as a pre dictor, are presented in the two rightmost columns of Table 7. It was found to be statistically significant, χ2(6) = 63.54, p < 0.001, suggesting that one can identify the probability to succeed in the courses based on the following parameters that were found to be significant: The SCOP at the 8th week (p < 0.01) and having an advanced diploma (p < 0.05). Again, the course name was found to be insignificant in this model. The model correctly classifies 66% of the cases (see the model evaluation), and similar to the model A above, the Hosmer-Lemeshow Goodness-ofFit tests (Hosmer & Lemesbow, 1980) found it to be well fitted. Both our models indicate that the early prediction models, based on students’ data collected before the course’s mid-point, enable identifying students who will probably succeed as well as those who probably will not succeed and might need extra attention. 6.2.3.2. Evaluation of the models. In each model, we defined a student with a probability of 0.5 or higher to succeed as someone who will probably successfully complete the course, and a student with a probability of below 0.5 as someone who will probably not complete the course. The overall correct predictions of model A are higher than those of model B. This means that model A is more accurate than model B and that the submission of assignments is a strongerpredictor than the video sessions’ opening patterns. We further evaluated the models by plotting the area under the curve (AUC) to estimate their accuracy based on the Receiver Operating Characteristic (ROC) curve (See Tables 10 and 11 for details). The ROC curve is plotted with sensitivity in the Y-axis and specificity values in the X-axis. The sensitivity measures the probability that a given statistic correctly predicts the actual condition with respect to a pre-defined threshold. tails). To check the robustness of the logistic regression results, we ran a forward entry stepwise logistic regression. In this process, one starts with a blank slate, and all variables are assessed for their potential predictive power (Li & Liu, 2019; Osborne, 2015; see Table 8). In a forward entry, one starts with a blank slate, and all variables are assessed for their potential predictive power. The variable with the single greatest relationship is entered into the equation, and then all remaining variables are assessed according to whether they add significant predictive power to the equation above that variable in the equation. The next strongest predictor is added, and the process continues until no variable

6.2.4. Decision tree algorithm and model compression

As explained above, the multicollinearity between the optional assignments’ submission status and the SCOP forced us to run separate logistic regression models for each variable. To test these variables together in a single model that predicts students’ success in the course, we utilized a DT algorithm, as detailed in the Methods section. Results are presented in Table 12. The dependent variable was the course completion status. We ran a separate DT analysis for each week of the course and updated the values of the SCOP and the assignments’ submission rate accordingly. The suggested independent variables were the SCOP of each week, the optional assignment submission status (starting from week 5), gender, the district of residence, and the most advanced diploma. The independent variables were selected for each week according to the CHAID algorithm (Lin & Fan, 2019). Fig. 6 exemplifies the DT procedure at week 8. The algorithm’s accuracy was found to be 76%using tenfold cross-validation. The model starts with a ‘parent’ node (node 0) containing all 797 students, which displays the number and percentages of students who completed or did not complete the course. From this point, the tree splits in the order of importance. The first optional assignment was the most significant factor regarding course completion. Thus, the ‘parent’ node splits into two ‘child’ nodes, one containing students who submitted the assignment (Node 1, left branch) and the other containing those students who did not submit it (Node 2, right branch). Next, each node splits by the SCOP value at week 8 into three ‘grandchild’ nodes.

Regarding the video recordings, we found that only a few interviewees attended the live sessions, and that the rest watched the recordings at their convenience. In addition, students reported that they communicated with each other through a social media platform (students’ WhatsApp group), which is external to the course. This platform allowed students to consult with each other and to answer questions. This finding supports previous studies (Rap & Blonder, 2016; Rap & Blonder, 2017) that found that students use social media platforms to interact with each other and discuss the course materials. Finally, we learned that the submissions of the assignments, both mandatory and optional, were also used as a learning strategy and for self-evaluation. These learning choices guided us in choosing the learning variables that could be used to construct a model to predict students’ success in the courses, namely, video sessions’ opening and the submission of optional assignments. This helped us develop predictive models and address the second research question (Q2): “What indicators of online learning can predict course success, and at what stage? Based on Model A, we found that already at week 5, in accordance with the deadline for submitting the first optional assignment in the studied courses, we could identify students who had a high probability to successfully complete the course. This finding shows that an optional assignment, which we view as an indication of student choice, is an important predictor of course completion. This expands on previous studies that found that the more assignments students completed on time and the earlier they did so, the better they performed on the quizzes and the final exams (Baker et al., 2020). Model B showed that by week 8, around the midpoint of the semester, students who eventually successfully completed the course had different accumulated video opening patterns than those who did not succeed in the course. The SCOP variable, which is the main predictor in this model, is an indicator of students’ time management, since it reflects their advancement in the course from week to week. Naturally, each model’s specific week in the course can differ between courses and institutions. Nevertheless, our models show that students’ choices (whether to submit an optional assignment or open the video sessions) are key predictors of students’ likelihood to successfully complete the course. Comparing the two logistic regression models, we found that except for the last two weeks, Model A was a stronger predictor of course success than was Model B. This can be understood considering that submitting an assignment better represents active learning than opening a video (Gabbay, Cohen, & Festinger, 2020; Glick, Cohen, & Gabbay, 2020). By active learning, we mean that participants are dynamically or experientially involved in the learning process, which is known to be a more important feature of successful online learning (Davis, Chen, Hauff, & Houben, 2018). Future research should examine whether embedding active learning features within the video sessions can increase the predictive power of video opening variables such as the SCOP used here. Improved prediction for students who didn’t complete the course was achieved with the DTCHAID model.

Concluzii:

Nevertheless, we recommend that new dashboards for a specific course be created along with an evaluation process for the relevant courses, and that it will involve both researchers and course staff. This evaluation would provide guidance regarding choosing the most relevant SRL and learner engagement indicators. Academic institutions could also consider embedding an automatic weekly statistical analysis in a dashboard that will present the lecturers with the probability of students’ success in the course. Finally, the findings from this research are especially relevant in the context of the covid-19 pandemic and will continue to be important in the post-covid-19 world. Online learning has been growing considerably in recent years and became extremely important with the rapid transition to online learning in both schools and universities following the coronavirus outbreak. This dramatic change continues to attract the attention of educators, researchers, policymakers, and the media to various learning theories that can promote more effective online learning, as well as addresses the major open questions in the field regarding students’ persistence, the learning process, course design, and student-instructor interactions.

Probleme ramase nerezolvate:

Another limitation of this study stems from ethical considerations that restricted us from connecting the online activities of individual students to their interview data. Thus, we could not successfully relate to an online learning pattern in the course on a personal basis, and it remained in the framework of statistical analysis. Finally, an important limitation is related to interpretation of a repeated video opening. We knew from the interviews that many students re-watched the course’s videos, but we could not accurately assess their re-watch patterns. This was because our data on multiple video opening included both students who re-watched lessons and those that simply re-opened them due to technical issues. Therefore, we decided not to analyze this pattern using the analysis.

[8]

Scop:

In this paper, we propose a novel stacking ensemble based on a hybrid of Random Forest (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosting (GB), and Feed-forward Neural Networks (FNN) to predict student’s dropout in university classes.

Abordare:

To address the issues presented in the literature,this study proposes a novel stacking ensemble based on a hybrid of Random Forest (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosting (GB), and Feed-forward Neural Networks (FNN) to predict student’s dropout in university classes.

Decision trees and rule-based classifiers, on the other hand, are white-box models that are moreunderstandable and easily interpretable because they expose the reasoning process underlying the predictions. Clustering algorithms or association rule mining are other options. Correlation analysis of course grades and attributes defining credits obtained by students and their average grades can also be useful (Lang et al., 2017).

The raw dataset used to conduct experiments in this study was collected directly from Constantine the Philosopher University in Nitra records from 2016 to 2020 (Kabathova & Drlik, 2021). The initial data contains 261 samples and 12 features of students registered in the preparatory lesson on database systems. The features in the raw datasets include information about [access], [tests], [tests\_grade], [exam], [project], [project\_grade], [assignments], [result\_points], [result\_grade], [graduate], [year] and [acad\_year]. The output variable has two values either 1 for non-dropout or 0 for studentswho dropped out of the university course. Among all students, 210 (80.46%) passed the course successfully and have not dropped out the school and 51 (19.54) failed to pass. In order to find the screened features for model building, the dataset cleaning process was performed to deal with, irrelevant, noisy, and inconsistent data (M ́arquez-Vera et al., 2016 ). The null and unre alistic values were dropped. Additionally, for categorical features, one-hot encoding was performed to transform integers to binary vectors. Besides, the feature selection process was conducted to decide the input variables for model building. As shown in Fig. 1, a correlation heat map was designed to analyze the correlations between input features and output variables. There exist a medium correlation between important features and grade points. Features such as tests, access, and project have a strong correlation with grade points. As a result, strongly correlated features are considered in the model building due to their highest impact on student outcomes. In the last stage of data preprocessing, the dataset was normalized using a standard scaler to eliminate the mean and scale it to unit variance (Obonya & Kapusta, 2018).

4.Methodology

In this study, a novel stacking ensemble made up of a hybrid of RF, XGBoost, GB, and FNN is proposed to predict student’s dropout in university classes. A hybrid of four models is proposed to make a powerful meta-learner (Xing et al., 2016). Stacking generalization is defined as an ensemble modeling technique to merge several classification models via a meta-classifier (Wolpert, 1992). The process of combining multiple classification models applies non-linear weightings for low-level predictors to minimize the generalization error rate and enhance the prediction accuracy (Wolpert, 1992). The proposed approach consists of two layers. In the first layer, temporal predictions of the RF, XGBoost, and GB are generated using a complete training dataset to extract the merits of each base classifier. In the second layer, the predictions generated in the first layerare fed to the FNNmodel to compute the final prediction of student dropout using cross-validation (Jiang et al., 2020). As shown in Fig. 2, the proposed approach has four important stages namely, feature engineering and selection with rationale, dataset splitting, final prediction, and evaluation. In the proposed stacking ensemble, the raw data with messy and irregular features will be processed through multiple classification models, and valid features will be extracted. Stacking’s learning ability stems primarily from the representation of features, which is consistent with the structure of neural networks (NN). The first layer in Stacking is analogous to the first N-1 layer in a NN, while the second layer in stacking is analogous to the last output layer in a NN. Stacking’s first layer can be thought of as a highly complex nonlinear feature converter. In stacking, different classifiers represent heterogeneity for different features. To effectively extract features from raw data, the first layer’s base classifiers must meet two requirements such as high accuracy and high diversity. RF, XGBoost, and GB are chosenas the first layer’s base models in this study. All base classifiers accomplish learning tasks by combining multiple learners, but their modeling concepts are completely different. These three base models were chosen and combined in the first layer of the proposed model because of their similarities and differences, and they all performed well in crossvalidation as well as returning the best accuracy.

Because features are extracted using complex non-linear transformations in the second layer of our model, complex classifiers in the output layer are unnecessary. FNN is a good candidate because of its simple structure and additional benefits.

1.1. Artificial neural network (ANN)

ANN is a commonly used model for prediction systems and it maps several inputs data into a set of suitableoutputs (Thanh et al., 2019). The ANN is widely used by many researchers due to its ability of parallel computation, ease of implementation, and swift operation. This model contains a set of many algorithms in which the intelligence of human beings is integrated into the computing machines that can solve extremely complex problems with excellent accuracy performance (Sun et al., 2016). ANN structure consists of three layers, namely, the input layer which functions as raw data receiver, the hidden layer for which the nodes of one layer and next layer are connected entirely, and the output layer which functions as a result displayer (Shahin et al., 2008). The process of learning originates from the error backpropagation utilizing the gradient descent research approach while the predicted output is represented by the symbolic function ̂y denoted by Eq. (1) (He et al., 2009)

4.1.2. Gradient boosting (GB) model

GB or gradient boosting tree is an ensemble ML method utilized to solve both classification and regression problems (Ikeagwuani, 2021). The GB was first introduced by (Friedman, 2001) as multiple additive trees that can be usedto improve the decision tree approach by utilizing stochastic gradient boosting. The main objective of GB is to decrease a loss function by stirring on the opposite side of the gradient (Friedman, 2001). This means that the GB builds the new base learners to principally correlate with the negative gradient of the loss function in the prediction process. A loss functionis an indicatorof how gooda model is in the prediction process given a number of different parameters. Generally, when the loss functionis small, the performance of the model becomes accurate and efficient (Friedman, 2001). The input training samples are represented as {(x1, y1), (x2, y2), ..., (xn, yn)} while the output is the probability of predicting one among the target classes (crash injury severity levels). The detailed description of this algorithm is summarized in Algorithm 1. In the GB model, the gradient descent diminishes complex loss functions that cannot be reduced directly. The loss function to decrease is denoted as L. Initialize the model with a unique forecast value F0(x) with the average of training target values. For initial iteration m = 1, calculate the gradient of L in relationship to prediction value F1(x) and then fit a base learner to the gradient constituents. Compute the magnitude multipliers and update the function to get the predictionvalue. The process continues recursively until the final result sign(Fm(x)) from the collection of all regression trees produced throughout the iteration is computed. The GB employs cross-validation and Out-of-bag (OOB) approaches to find out the optimal number of boosting iterations. OOB permits on-the-fly calculation with no need for persistent model fitting (Friedman, 2001).

4.1.3. Random forest

Random forest (RF) is a ML algorithm that was first invented by Tin Kam Ho (Yuan & Hu, 2016). It is used to train weak learners so that identical problems get solved and combine them to obtain accurate results (Liaw & Wiener, 2002). RF was extendedby (Breiman, 2001) as an accurate classifier made by a group of tree predictors in such a way that there exists the dependence of each tree on independent values of a sampled random vector. Fig. 4 explains the general idea of the RF procedure. It is made up of four important parts: Firstly the training data sample is selected from a given dataset with the help of bootstrapZ times (Breiman, 2001). From this process, each sample set has an equalchance of being selected to represent the training set in every round the sample is selected. In the second phase, the decision trees are formed through the process represented as tree-based learner generation. In this phase, the node splitting process is responsible to select randomly the features to represent the set of tree-basedpredictors. The third phase consists of a forecasting process where the set of selected features is used to present the outcome of each tree predictor. Finally, the outcomes from each tree-based learner are merged where any predictor has equal proportion to the final result.

4.1.4. Extreme Gradient Boosting (XGBoost)

The XGBoost model is one of the most commonly used algorithms introduced by (Chen & Guestrin, 2016) to solve prediction problems. The objective function of XGBoost relies on regularization to the cost function terms such as tree depth and leaf nodes ’ weights (Zhu et al., 2021). This means that this model has the capacity of enhancing the performance of building trees when the iteration process reduces. Regularization to the cost function which makes XGBoost a regularized boosting technique is mathematically denoted by (Chen & Guestrin, 2016):

5.1. Model training and hyper-parameters tuning

aximum depth of the tree (Chen et al., 2015). In the second layer of our method, the input features were trained using FNN. The input layer of FNN consists of five neurons, each representing one input variable, while the output layer consists of one neuron which represents the output variable (Eldan & Shamir, 2016). To determine the number of hidden layers and the number of neurons in each layer, a replication searching process was followed to find out an optimized method with accurate prediction performance (Pontes et al., 2016). After performing several iterations, the optimal topology which produced efficient prediction results was found to be two hidden layers comprising of five neuronswith tangent sigmoid activation function and two neurons with Softmax activation (Pontes et al., 2016). The leading training function which resulted in the best prediction accuracy in the FNN is the Levenberg–Marquardt (Yu & Wilamowski, 2018). This was selected after comparing with other training algorithms such as Bayesian regularization backpropagation, scaled conjugate gradient, variable learning rate backpropagation, resilient backpropagation, and BFGS quasi-Newton (Yu & Wilamowski, 2018). After optimizing the parameters for the models in the first layer and the remaining in the second layer, a novel stacking ensemble model proposed was tested using the same testing set, and finally, the prediction performance was assessed based on the contingency table.

The discussion of the various performance metrics confirms that the selected binary classification models can be used to predict students’ dropout or Non-dropout at the individual course level, even when the dataset is scarce and has a limited number of input features. However, before the classifiers can be used to predict student dropouts, a broader set of performance metrics must be investigated. This statement is consistent with the findings of other research papers published in the domains of AI and ML.

Concluzii:

On the dataset collected from 2016 to 2020 at Constantine the Philosopher University in Nitra, the proposed method has demonstrated greater performance when compared with the base models using testing accuracy and the area under the curve (AUC) evaluation metrics under the same conditions. Based on the findings of this study, students at the risk of dropping out the school can be identified based on influential factors and different agents of education can refer to this infor mation for early intervention in the uncontrolled behavior that can lead to the risk of dropping out and take proactive precautionary measures before the issue arise.

The study presented here helps to solve the problem of student dropouts at the course level. The results demonstrated that, despite a small dataset, appropriately selected indicators that do not require access to system logs can be beneficial if different performance metrics are evaluated. The predictive models were fed with data gathered about students’ online learning environment activities and partial achievements. Simultaneously, a proposed methodology is reliable for predicting course completion when there is enough time for educators to intervene timely. The results obtained for our method can help in reducing the dropout rate based on identified students that are likely to be affected and the influential factors. Moreover, once the students who are at the risk of being affected are identified, the agents of education can gather their forces to take efficient measures of eradicating the issue. Furthermore, the different learner’s motivation strategies can be emphasized to improve the performance and help learners finish their programs successfully. For future research, other computation approaches such as deep learning and other hybrid models can be used to predict student dropout and compare the results with the findings of this study. Other influential factors not presented in this study must also be considered and a further feature analysis study is recommended to help the agents of education in handling successfully the issues of student dropout.

Probleme ramase nerezolvate:

Another limitation of this study is that different runs of the courses provided different data to be analyzed. As a result, it was difficult to determine which attributes are important enough to predict the stu dent’s performance in general. The next limitation is the selection of the classifiers used in this case study. Literature reviews of the current state of the researchand trends in LA and EDMprovide numerous examples of more or less advanced ML techniques that can be used to predict early student dropout at the course level. However, none of them have produced noticeably better results thus far. Because the main goal of this study was not to find the best one, the final choice of the classifier used in this case study allows mentioning that good prediction could also be achieved using stacking ensemble, but the performance metrics must be evaluated. The case study’s final flaw is a type of appropriate intervention that is not discussed in depth. The case study has already demonstrated that the student’s dropout at the university level can be predicted based on the activities chosen. The reason for this unflattering state, on the other hand, remains unknown and necessitates further investigation. It would be interesting to investigate whether other types of activities have a similar impact on student engagement and which types of activities can be exchanged during the intervention phase.

[9]

Scop:

This document presents an experimental study to obtain a predictive model that allows anticipating a university dropout.

This study aims to find a model based on computational learning algorithms (decision trees and neuronal networks) that anticipates university desertion aimed at reducing the desertion rate in degree programs in Engineering, Social and Administrative Sciences.

Abordare:

The study uses 51,497 instances with 26 attributes obtained from social sciences, administrative sciences, and engineering collected from 2010 to 2019. Artificial neural networks and decision trees were implemented as classification algorithms, and also, algorithms of attribute selection and resampling methods were used to balance the main class.

This experimental study was based on educational data mining methodologies and computational learning algorithms such as neural networks and decision trees.

The decision tree algorithm C5.0 (in its commercial version known as C4.5) is an extension of ID3. It can work with continuous values for the attributes, separating the possible results into two branches. The trees it generates are less leafy because each leaf does not cover a particular class but a class distribution. C5.0 forms a decision tree from the data employing recursively executed partitions, according to the depth-first strategy. Before each data partition, the algorithm considers all possible tests that can divide the data set and select the test results in the highest information gain or the highest information gain ratio. For each discrete attribute, a test with n results is considered, where n is the number of possible values the attribute can take [33].

Fig. 1 shows the method used in this study, which comprises two stages: 1) data processing and 2) classification model. The first stage bases on the knowledge discovery process (KDD), starting from describing the data set, processing, and model construction, described in points A and B of this section. In the second stage, ranking methods are applied to find the descriptive attributes. The decision trees algorithm and neural networks are applied to obtain classification models. Finally, the classification model with the best performance is obtained; this process is described in the results section.

1. Dataset Description

Table III presents the distribution of data used in this study, obtained from three deaneries distributed among 25 programs collected from fall 2010 to fall 2019 (Table III).

1. KDD Process

The selection of attributes utilizing algorithms allows improving the input data’s quality with the elimination of attributes that are not relevant [41]. In this study, we use subsection selection with CFS (Correlation Feature Selection), the filter method (Chi-square and gain information), and wrapping (Random Forest).

Data mining (classification model): The methods: X^2, Relief, and SOAP (Selection of Attributes by Projections) are applied since they increase accuracy, decrease overtraining since they eliminate data with better significance, and increase training speed [28]. The C5 decision tree algorithm and the artificial neural network algorithm with multilayer perceptron with 1 and 2 layer topologies with different numbers of neurons also obtain the best classification model obtained between both. For the evaluation of each algorithm’s performance, the metrics of precision, accuracy, specificity, and sensitivity were used, as well as the balanced accuracy and the Matthew’s Correlation Coefficient (MCC). For the validation of the models, cross-validation with base ten is employed. In the training part, the balanced subset of data applies. The proportion for the Deserter(S) class is 34.5%.

1. Training and Testing

Executions perform with different proportions in several instances; the following strategies are applied: 1) training with 70%, 80%, and 90% of the available data and 2) For the initial evaluations of the models, base ten cross-validations use.

1. Feature Selection

Two methods for feature selection are used to find the most relevant attributes. To determine the order of importance of the attributes rank methods (chi\_square and gain information) were used and filter methods were used to determine a subset with the most significant attributes.

Concluzii:

A. Features Selection

Two feature selection methods are applied to identify the most relevant attributes. The first method, feature selection algorithms, was used with rank methods (chi\_square and gain information) to determine the most relevant attributes. In which it is observed that similar results were obtained with both methods (Fig. 2). A second method was performed with filter methods to determine a subset with the most significant attributes—also, the wrapping algorithm applies with the random forest (Table V).

Fig. 3 shows the order of attributes from least to most important with box diagrams, thus showing the dispersion of data for each. It can be noted that the variable FORANEO is not significant, and the variable REPROBADASP is significant. That means that the most significant variables are related to the failure rate, absences, semesters completed, age, average income, current average, and progress according to credits. This analysis also allowed us to detect some inconsistencies in the attributes PROM\_PANTE, AVANCE\_XCRED, and FALTASP:

* PROM\_PANTE.

Most of them represent zeros values, which is not necessarily incorrect since these variables are affected by new entries and re-entries. When they are new entries, it is expected that they do not have a previous average because there is no history. The same happens with their advance by credits and faults. However, atypical data is found, where it is not typical for values greater than 0 and less than 6 to exist, since it indicates that someone who failed the previous semester and registered for the next period without credit. Thus, 90 records are discarded due to the inconsistency described.

• FALTASP.

Five cases were eliminated from the data set with absences greater than 500, which could indicate an inadequate capture in the system.

• ADVANCE\_XCRED.

2 cases with incorrectly calculated data from the system were discarded.

The next phase was to determine which attributes are applied for model training. For this task, decision trees with base-10 crossvalidation is used as a classification algorithm. Fig. 4 shows the results with three training ratios, 70%, 80%, and 90%, and with attributes according to the methods of Correlation (CFS), Consistency, and Wrap (random-forest). It can see that the best performances were obtained with a training ratio of 90% and 80% with selection methods by Consistency and Wrap-RF. The selection of attributes by CFS causes the model to lose sensitivity with any training ratio, being the least desired option. However, it highlights that it reduces up to 9 significant attributes, which could be useful when training time must be optimized at the expense of loss of fit.

From these results (Fig. 4), it determines that the most important attributes are those obtained by the wrapper (Random Forest) due to the performance obtained. These are ASIG\_INSCRITOSP, FALTASP, AVANCE\_XCRED, PROMEDIO\_AC, PROM\_PANTE, REPROBADAS3P, CRED\_PANTE, PROM\_INI, AGE, REPROBADAS2P, FALTAS\_PANTE, SEMESTRE\_PANTE, REGISTERED\_PERIODS, REPRODUCED1P, SUPPORT, PANT\_PLAY, REPRODUCED1X, AUTUMN, SPRING, REPRODUCED2X, FEMALE\_GEN, MALE\_GEN, NINGROS, REPRODUCED3X, and STADIUM.

B. Classifier Model

With Random Forest Decision Trees The decision tree algorithm C5.0 applies balanced data, and the attributes of Table IV. Fig. 4 confirms that the attributes selected by RF-Wrap give the best results, although they are very similar to those obtained by Consistency. The lowest evaluation is obtained using only the attributes by CFS. It can be seen that only the accuracy metric is lower (~88%). This metric indicates the proportion of true positives of those classified as dropouts. The model’s evaluation improvement, the balanced accuracy, and Matthew’s correlation coefficient were obtained. The balanced accuracy is maintained for the attributes obtained by Consistency and RF envelope; Matthew’s correlation coefficient indicates a robust positive relationship above 70%. In this way, the list of the most relevant attributes shown in Fig. 5 is obtained.

C. Classifier Model With Artificial Neural Network

Fig. 6 shows that the attributes obtained by CFS have the lowest values. However, as attributes are added, there is an improvement in the results. Although accuracy is high, sensitivity is low, meaning that it can detect about 57% of defectors (depending on network configuration and several attributes). On the other hand, it is analyzed the result with the Matthews correlation coefficient (MCC) and balanced accuracy.

D. The Best Classifier Model

According to the classifier and configuration used, Fig. 8 shows the best results of artificial neural networks and decision trees. For the case of decision trees, the best results are given by attributes taken by RF-Wrap. For Neural Networks, the most appropriate configuration was a hidden layer with 12 neurons. The results show that decision trees are the best classification algorithm for detecting deserters.

Thus, the Table VI presents a comparison of the application of neural networks. There is a difference between the number of records and variables studied in each study, determining factors in obtaining performance metrics such as accuracy and sensitivity. Although in [15], [16] the accuracy was higher than 96%, while in our study, it was 76.79%, this does not mean that neural networks are not adequate to predict dropout university in our case. Nevertheless, they are a function of the type of data, variables used, classification methods used in the neural networks, and the attribute selection methods applied before the classification, as mentioned in the explanatory features section.

The Table VII shows a comparison of the results obtained in our study with those obtained in the literature. It can be observed that the difference between accuracy and sensitivity is notorious and the number of variables and data analyzed. As in neural networks, these results’ difference lies in the data set analyzed and the attributes. In summary, the closer the percentage of metrics is to 100%, the better the classifier.

In the first analysis of attributes, some that do not seem so relevant can be discarded, such as failed subjects, period (autumn or spring), and gender. However, comprehensive analysis with attribute selection by envelope - random forest shows that the models’ maximum performance is obtained using most of the attributes. On the other hand, the class balance allowed us to improve the performance metrics of both algorithms. Mathew’s correlation coefficient and balanced accuracy provided a better evaluation of the models, allowing the results to be unaffected by variations in accuracy, precision, specificity, and sensitivity metrics. Decision trees obtained the best Matthew correlation coefficient of 87.43% and balanced accuracy of 94.34%. On the other hand, the tests performed indicate that increasing the number of neurons in the hidden layer of ANN could improve performance. However, it requires more processing power since server training can be more than an hour-long, and it is the main reason for not having performed more tests. Since the best model has been the one obtained by decision trees, it has been implemented in the Enterprise Resource Planning (ERP) institutional system, which helps the student to receive the accompaniment he requires so as not to interrupt his studies; this effort translates into monitoring about 450 students per period belonging to the deaneries of social sciences, administration, and engineering. In this way, the student monitoring area improves its administrative tasks using the model because it obtains a list in a matter of seconds and avoids consolidating a report of several Excel documents, even from different areas (admissions and school control).

The results show that the best performing model was that of Random Forest with a Matthew correlation coefficient of 87.43% against 53.39% obtained by artificial neural networks and 94.34% accuracy by Random Forest. The model has allowed predicting an approximate number of possible dropouts per period, contributing to the involved instances in preventing or reducing dropout in higher education.

Probleme ramase nerezolvate:

To continue with the research, we have considered using additional attributes such as payment history, debts, campus access, and similar, as well as implementing stratified sampling by deanery for class balancing in the model’s training and dividing data into new entries and re-entry. Also, we intend to use other classification algorithms such as near neighbors, vector support machines, logistic regression, and use a combination of classifiers to generate a more robust solution. Also, we plan to use cloud services such as Machine Learning in AWS, Azure Machine Learning, BigML, IBM Watson, TensorFlow, or some other computer learning solution to improve processing times and, finally, deploy models to classify online mode.

[10]

Scop:

This paper proposes the student performance prediction using CNN (Convolution Neural Network) and BPSO (Binary Particle Swarm Optimization) based feature selection method. In this study, classifiers are made for 2-class and 5class predictions.

Abordare:

In this context, this study aims to predict student performance; for this, a common dataset from the UCI machine learning repository is utilized, with this it was possible to compare techniques already consolidated in the scope of the EDM, with the machine learning technique. The approach used was supervised learning for classification (Ajibade et al., 2018), in which students’ grades are predicted, but these were divided into four categories, their numerical values not being used. With these predictions, it was also possible to verify if the attributes that make up the database are sufficient to generate effective models in predicting student performance, in addition to evaluating BPSO-based feature selection to ensure redundancy management and proper training in supervised classification. Finally, this study intends to make available to those interested in the area, a document that presents in detail how to carry out the educational data mining process.

Feature extraction (Zaffar et al., 2020) is an important step in classification because the effectiveness of a learning model depends on input variables (substantial features) that describe student characteristics and can be used to predict student performance. This data refers to the performance of secondary school students in two Portuguese schools. Attribute data (including student grades, demographic, social, and academic characteristics) were collected using school reports and questionnaires. Two sets of performance data are provided in two different subjects: Mathematics (Math) and Portuguese (Po). In Athani et al. (2017), both datasets were modeled using binary/five-level classification and regression assignment. Important note: The target G3 attribute has a strong correlation with the G2 and G1 attributes. This is because G3 is the final value (issued in the 3rd period) and G1 and G2 are the values of the 1st and 2nd periods. It is more difficult to predict G3 without G2 and G1, but these predictions are much more useful.

3.2. Attribute selection by BPSO Binary particle swarm optimization (Kumar and Bharti, 2019) is used to select a subset of M features from a set of N features of a database, (where M<N), in order to reduce redundancy in the database so that optimal results can be achieved. BPSO finds optimal results from the search space of the candidate solution using the equation below: 𝑣′ = 𝑤 ∗ 𝑣 + 𝑐1𝑟1(𝑥𝑃𝑏𝑒𝑠𝑡 − 𝑥(𝑡)) + 𝑐1𝑟1(𝑥𝐺𝑏𝑒𝑠𝑡 − 𝑥(𝑡)) (1) 𝑥(𝑡 + 1) = 𝑥(𝑡) + 𝑣′ (2) where, x(t + 1): New position of the particle in search space x(t): Current position of the particle in search space w: Intertia (assumed to be 0.8) v: Velocity of particle

c1, c2: Cognitive and social attraction coefficients (assumed to be 1.414) r1, r2: Random numbers ∈ (0,1) xPbest: Local best position xGbest: Global best position

3.2.1. Search space and particle coding Very first we need to create a search space where all possible candidate solutions can be defined. Let’s a feature has N attributes i.e., 𝐴 = {𝐴1𝐴2𝐴3 ... ... 𝐴𝑁−1𝐴𝑁} then we need to code each candidate solution to a binary string of length N. we represent each attribute by 1 or 0 to show its presence in the selected subset where the MSB of the binary string represents 𝐴1 and LSB represents 𝐴𝑁. To perform a valid training of classifier we need at least one attribute of data so the lower boundary of search space becomes 1 coded in the binary string of

length N. If all the attributes are selected (worst case) for training, in that case, the upper boundary of the search space becomes 2𝑁 coded in the binary string of length N. For example, if the database has a feature of 7 attributes 𝐴 = {𝐴1𝐴2𝐴3 ... ... 𝐴7} then N becomes 7 and the total possible subset are 27= 128. If a candidate solution is [0 0 1 0 0 1 1] then selected subset will be 𝑆 = {𝐴3𝐴6𝐴7}.

Concluzii:

A model was generated for the BPSO-CNN algorithm used in this study, from its application in the maths database, it should be noted that for the application of the algorithms all attributes and BPSO selected features were used in the database for 2 classes described in Table 1, and all attribute and optimized attributes used in the final grade classified in 5 categories: A, B, C, D, and F. The proposed methods give higher accuracy of 93.3 % for 2 classes whereas 5 classes prediction gives 86.21 %. A model was generated for the BPSO-CNN algorithm used in this study, from its application in the Portuguese database, it should be noted that for the application of the algorithms all attributes and BPSO-selected features were used in the database for 2 classes described in Table 1, and all attribute and optimized attributes used in the final grade classified in 5 categories: A, B, C, D, and F. Various results of the proposed method are provided in Tables 2-6. The proposed methods give higher accuracy of 96.67% for the 2 classes whereas 5 classes prediction gives 86.11 %. Table 7 and Table 8 present a comparative analysis of the proposed work with previous work for UCI Maths and UCI Portuguese datasets respectively.

To compare the results obtained with the methods proposed in this work with the results obtained in other works, tests were carried out with the methods of classification used in some works. This comparative study was carried out as described. The proposed BPSO-CNN method for Mathematics and Portuguese datasets outperforms traditional CNN with an improvement of 1.6 % in accuracy. Whereas the Logistic Regression based method gives 62.05% and 67.69% in mathematics and Portuguese dataset respectively, which is very low compared to the proposed method.

The main objective of this study was to predict the performance of students using a public dataset and compare the prediction effectiveness of the model generated by CNN. A hybrid structure of CNNBPSO was the most accurate. With that, it was possible to verify that a database composed of records of notes, frequency, and characteristics demographics, (social and school) is sufficient to carry out the generation of effective models in the prediction of student performance.

Furthermore, it was possible to identify that the attributes referring to the student grades and absences are more predictive of performance than student records. Regarding the performance prediction, the proposed EDM technique was adequate, in which the results achieved are the following: CNN with an accuracy of 91.14% in 2 classes of Maths data, BPSO optimized features with 93.33%; Where 5 classes Maths with 84.40%; CNNBPSO with 86.21%. Similarly, CNN with an accuracy of 93.85% in 2 classes of Portuguese data, BPSO optimized features with 96.6%; where 5 classes Portuguese with 85.59%; CNN-BPSO with 86.11%. Regarding the set of attributes with the greatest influence on the prediction of student performance, it was identified as a relevant finding.

The proposed system claims an outperforming accuracy of 96.6% with various previous research works as well as found that the majority of attributes related to school activities as compared to data on demographic and socioeconomic characteristics.

Probleme ramase nerezolvate:

[11]

Scop:

The main key features of this study examine the status of students (i.e., pass or fail), grades, and final exam marks resembling the debilitation in their performances and achievements. Underpinning the aims of the present study is the notion that the research described sheds light on the classification in conjunction with ensemble methods evaluated on a new collection of features, including demographic, academic, personal, family, psychometric, and learning log. further, we devised attribute selection which is based upon both a filter and a wrapper-based method. the main contribution of this study is the exploration of the best impactful features that play a vital role in predicting students' performance. Second, analysis of attribute selection based on improvised filter and wrapper based method. Third, a novel ensemble classification approach extracts invisible and intrinsic connections between the feature of students and their performances. fourth, the prediction of students' performances shows accurate and optimized outcomes for the betterment of education institutes.

Abordare:

Thereby, our analysis induced with the dataset that provides one of the features known as learning logs. it enables students to log into the system before final exams for their assessment and perform activities like questioning, clarification, and interpretation [8]. in addition, the teacher construes the learning session that shows the superiority of the contribution is imitated in the students' participation. we used quantitative and qualitative indicators before exams, such as the learning logs and psychometric features, in conjunction with Data Mining (DM) techniques to determine and build substitute representation and models for underlying data. in contrast, we use educational data in predicting the academic performance of students that in turn evaluating the effect of the different features, namely Demographic (DE), Academic (AC), Personal (PE), Psychometric (PS), Family (FA), and Learning Logs (ll). we examined our analysis in the realm of Machine Learning (ML), namely, tree, logistic, function, neural network, rule, and instance-based algorithms with ensemble methods such as boosting and bagging, as they enhance the model's accuracy.

2. Data Description

The progression period was analyzed from 2016-to 2018, wherein each year, students were logged into the learning log session once before the final exams. as mentioned in section 2, the dataset is collected from the online educational system Smart Learning Partner (SLP) at advanced innovation center for future education, beijing normal university [5]. the data gathering consists of 11814 students from 31 local schools in beijing that were categorized into seven subjects. the features are divided into six categories: de, pe, ac, ps, fa, and ll. overall, students cover only three concepts in biology subject. the final grade comprises three partial components scores (i.e., PS, Learning Logs Score (LS), and Final Score (FS)). each score is weighted differently: ps is 25%, ll is 35%, and fs is 40%. This formula applies equally to all students and is a curricular definition fed into the data preprocessing (see Figure 1).

3. Methodology

ECA follows the classification models evaluated based on cross-validation and split validation (see Figure 2). in the pre-processing phase, raw data is transformed into an organized format, and it is collected from SLP into a single repository (i.e., a data management system). redundancy is a common problem occurred when integrating data, and this is why we used a centralized database system to fetch students' data uniquely via queries. data consistency is handled by filtering the missing and noisy data, and the dataset occupied by this study does not contain any missing and outliers.

Further, we transformed specified data from numerical values into bi-nominal values, such as attributes including 'live campus', and 'is\_only\_child' are converted into bi-nominal. The target variable (class label) is known as 'situation', which describes whether a student passes or fails. once data is ready, Attribute Selection (AS) is implemented using filter and wrapped-based methods. filter-based includes 'attribute evaluator' via search methods, namely 'ranker' and 'greedy stepwise'. on the contrary, wrapper methods use combinations of each pair of attributes and evaluate the best features that achieve higher accuracy. a 10-fold cross-validation model is performed in the model evaluation phase to validate the classification results. After that, ensemble-based classification is performed via split validation (i.e., 70% training and 30% testing). Finally, the comparison and results of ensemble meta-based classification against baseline model performances (via crossvalidation) are validated.

4. Experiment Results

The proposed framework evaluates the analysis of the classification models with ensembles methods that can achieve better performances in predicting students' performances. we explain three essential issues:

1. Attribute analysis using attribute selection methods,

2. Analysis of state-of-the-art prediction models

3. Prediction model performances via ensemble methods. experiments were performed using Rapid miner studio, microsoft azure studio, and weka. the performances are measured in terms of accuracy, kappa, and F1-score via both 10-cross validation (i.e., baseline performances), and split validation (i.e., ensemble methods). experimental results highlight the RQs as discussed in section 1.

4.1. Filter Based Attribute Selection Via Ranking with Models Another way to get feature weights in filter-based methods is to use the model-based attribute selection. it provides a resultant weight vector that describes whether an attribute is essential for the learning algorithm. the concrete calculation scheme is different for all learners. after that, we selected those models that can compute the weight score of attributes and evaluate their effect on student performance. weight score is evaluated using the models, including:

1. Decision tree - Gini Index (DT-GI).

2. Decision tree - Information Gain (DT-IG).

3. Decision tree -Gini Ratio (DT-GR).

4. Random Forest (RF).

5. Deep Learning (DL).

6. Logistic Regression (LR).

7. Naïve Bayes (NB). each attribute can be distinguished in each model through a stacked bar via its unique pattern/color (see Figure 3). It is worth noting that most selection of attributes remained the same as they found using the rank-based method, adhering to a total of 18 attributes and estimated score of weight ranging between 0.01 and 0.60. the weight score is described as the number of times the model ranks each attribute. 4.2. Wrapper Based Attribute Selection Using A Possible Combination Of All Attributes Attributes used in this study are devised into six categories, and results performances are measured in terms of accuracy, kappa, and F1-score. first, we choose a single category of attributes (i.e., de, pe, fa, ac, ps, and ll), and the results indicate that demographic attributes gained low accuracy. at the same time, academic and family features achieved higher and constant scores in each model (see Figure 4). at the same time, remaining attributes' scores (e.g., pe, ps, ll) fluctuate from low or high scores due to the worst performances of some classification models. In the combination of two categories (such as fa, ps, pe and ll were given good performances as shown in Figure 5. additionally, fa+ll, ps+ll, and pe+ll gained highest score (.89, 0.9, 0.93), (0.86, 0.60, 0.91), and (0.9, 0.7, 0.93), respectively. In a set of three combination of categories, including family and learning log in combination with demographic, personal, and academic attributes showed better performances. results indicate that de+fa+ll, pe+fa+ll, and ac+fa+ll achieved the better scores (0.90, 0.74, 0.93), (0.90, 0.74, 0.93), and (0.90, 0.74, 0.93), respectively (see Figure 6). In a combination of four categories; de+pe+fa+ll, pe+fa+ac+ll, and ac+fa+ps+ll achieved improved scores (0.90, 0.77, 0.93), (0.90, 0.74, 0.93), and (0.90, 0.74, 0.93), respectively (see Figure 7).

4.3. Effect oF Attributes VIA Decision Trees

Figure 8 presents features' effects, including the academic, family, and psychometric categories. These attributes include “distance-to-downtown,” “students living on campus,” and teachers with bachelor's degrees, which impacts the final situation of students where the rate of teachers with bachelor's degrees per school is more significant. similarly, family attributes such as mother and father students impact their final situation as pass or fail. thus, it concludes that students can pass their exams with superior scores, including family, academic, and psychometric attributes. additionally, failed students could emphasize their academic, family, and psychometric scores that may help them pass the exams.

4. Ensemble Meta-Based Model

This experiment integrates classification models with ensemble meta based models (i.e., bagging and boosting and adaboost). only seven classification models were selected in this experiment whose performances were better using 10-fold crossvalidation as reported in [7]. ensemble meta models can optimize the performances of the classifiers and evaluate via accuracy, kappa, and F1-score. the results indicate that k-star and DT-IG achieved the highest scores, respectively (see Table 1). overall, the results performances of classification models DT-GI, DT-GR, RF-GI, RF-GR, and k-Star gained superior scores when ensemble with bagging and boosting. results validate that performance of each classification model gained high accuracy when used with all the ensemble meta-based model (i.e., bagging+Boosting+AdaBoost ). for the most part, k-star and DT-IG were also found better when integrated with ensemble meta-based predition model. therefore, these results confirm that the ensemble meta-based can boost the prediction model performances.

Concluzii:

We present a novel classification approach that performs feature selection via an improvised filter and wrapper-based method. We testify our experiments on a large dataset composed of new features in different domains. the proposed framework validates the results via both 10-cross validation and split validation. moreover, ensemble meta-based model identify students' performances, whether they pass or fail in their final exams. it resembles an indiscernible and intrinsic connection between the feature of students and in their performances. our findings indicate that ensemble meta-based methods are devised with a classification model that can help predict students' performances. the results also indicate that features including family, psychometric, and learning logs are more impactful and can be given much attention in any educational setting in predicting student performance.

Probleme ramase nerezolvate: