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

[1] Mediu de invatare online

Random Forest si RNA

Gini Importance (Mean Decreased in Impurity)

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).

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. Toate caracteristicile însumează un total de 32. OULAD este selectat, deoarece un număr suficient de rezultate au fost publicate folosind acest set de date până acum.

Ținta predicțiilor din această lucrare este performanța academică a studenților. În OULAD, această țintă este exprimată prin patru categorii diferite: Rezultate Excepționale, Promovare, Eșec și Retragere. Repartizarea celor 32.593 de studenți pe baza rezultatului final este următoarea: Rezultate Excepționale (3.024), Promovare (12.361), Eșec (7.052) și Retragere (10.156).

Pentru simplitate, se decide să se împartă cronologia cursurilor abordate de studenți 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.

Autorii raționalizează motivul pentru eliminarea etapei de inginerie a caracteristicilor din cadrul propus, urmând apoi 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 adăuga un pas flexibil de selectare/reducere a caracteristicilor la modelul propus.

În această lucrare, importanța caracteristicilor a fost calculată printr-un algoritm Gini Importance or Mean Decreased in Impurity. Un clasificator Random Forest este un set de arbori de decizie, iar acești copaci au ramuri împărțite în funcție de caracteristici. Algoritmul de importanță a caracteristicilor parcurge fiecare arbore de decizie și explorează fiecare ramură pentru a analiza îmbunătățirea modelului – scăderea impurităților pe baza divizării. Gradul de îmbunătățire a modelului pentru fiecare caracteristică este adăugat la scorul său de importanță, însumat în toți arborii și normalizat pentru a însuma 1. Pe lângă importanța caracteristicii, este posibil să se determine similitudinea dintre caracteristici prin corelația de rang. Asemănarea caracteristicilor poate ajuta la găsirea caracteristicilor apropiate unele de altele și să se decidă dacă sunt redundante.

Rețelele neuronale discriminatorii sunt potrivite pentru a prezice performanța academică a studenților. Primul model de predicție al studiului este o simplă rețea neuronală artificială. Random Forest este al doilea model de predicție din acest studiu.

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 Random Forest, cât și RNA-urile fără Feature Engineering 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 cursul ales.

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%.

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;

Probleme ramase nerezolvate:

Lucrările viitoare se pot 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 pot concentra în plus pe evitarea acestor informații demografice.

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] Fizic, Colegii de inginerie India

Decision Tree, Naïve Bayes, k-Nearest-Neighbor, Support Vector Machine și Logistic Regression

Chi-Square

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 Logistic Regression.

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 studenții cu performanțe academice slabe 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 bazată pe filtre.
2. Identificarea celui 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 la două colegii de inginerie diferite din India. Setul de date include informații cu privire la factorii demografici, academici, sociali și psihologici din trecut, cu 27 de atribute diferite.

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

1. Scopul principal al predicției performanței academice a studenților este de a oferi sprijin în timp util studenților cu performanțe. 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.
2. 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. Atributele cu o valoare p mai mică de 0,01 arată o corelație foarte semnificativă cu notele elevului.

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.

În studiul de față, au fost aplicate cinci modele unice de învățare automată supravegheată. 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).

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.

Concluzii:

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% pentru performanții slabi cu setul de date echilibrat formulat cu ajutorul tehnicii de reeșantionare SMOTE.

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] Fizic, Universitate+Scoli

Decision Tree, Logistic Regression, Naive Bayes, Random Forest, Multilayer Perceptron

SelectKBest, Chi-Square

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:

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

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. Studiul folosește tehnica SelectKBest, ce selectează primele k caracteristici cu cele mai mari valori de scor pe baza testului Chi-Square, pentru a compara rezultatele reale și cele prezise.

Pentru clasificare, s-au proiectat două modele diferite de ansamblu. Unul constă din Multilayer Perceptron (MLP), Random Forest (RF) și Decision Tree (DT), iar celălalt este format din RF, Logistic Regression (LR) și DT.

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%.

Prima 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, pentru o mai mare scalabilitate, autorii intenționează să aplice abordarea predictivă propusă la diferite seturi de date, să experimenteze diferite tehnici de selecție a caracteristicilor și să implementeze alternative pentru tehnica de clustering DB-Scan.

[4] Fizic, 2 scoli secundare portugheze,

Decision Tree, Random Forest, Naïve Bayes, Multistrat Perceptron, Jrip

Feature Subset Selection Wrapper

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.

Scopul principal al acestui studiu este de a ajuta profesorii să identifice elevii care sunt expuși riscului de eșec și să-i ajute să își îmbunătățească rezultatele educaționale.

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.

Abordare:

Acest studiu folosește două seturi de date disponibile în mod deschis pentru a prezice performanța elevilor. Pe lângă note, setul de date include informații despre demografia elevilor, statutul socioeconomic și școala pe care o urmează. Există 33 de variabile în total în ambele seturi de date.

Metoda Feature Subset Selection Wrapper a selectat cele mai relevante atribute și a crescut ratele de precizie. Selectarea atributelor 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, se folosesc filtre și wrappers pentru a alege caracteristicile finale. 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.

S-au propus inițial cinci algoritmi de Deep Learning, și anume Decision Tree, Random Forest, Naïve Bayes, Multistrat Perceptron și Jrip.

Concluzii:

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, cinci abordări binecunoscute de extragere a datelor, cum ar fi Decision Tree, JRip, Naive Bayes, Multilayer Perceptron și Random Forest au fost utilizate în experimente.

Rata de precizie a unui algoritm de Decision Tree a fost crescută de la 70,80% la 80,56% prin utilizarea abordării Feature Subset Selection Wrapper. 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%.

Ca rezultat, folosind metodele de extragere a datelor utilizate, succesul studenților ar putea fi prezis cu o acuratețe rezonabilă.

Probleme ramase nerezolvate:

Pentru a spori generalitatea metodelor, 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] Hibrid, Universitate Taiwan

Kruskal-Wallis H

Scop:

În acest studiu, se examinează 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 pentru studenții care participă la un curs de contabilitate al unei universități din Taiwan. Experimentul s-a desfășurat de la începutul semestrului până la examenul intermediar. Cursul a folosit BookRoll, un sistem de lectură a cărților în format electronic dezvoltat de Universitatea din Kyoto, unde instructorul încarcă materiale de studiu înainte de fiecare clasă, iar studenții pot efectua diverse activități în timpul lecturii. În acest curs, mediul de învățare online Moodle a fost folosit împreună cu predarea față în față. Activitățile Moodle au inclus accesarea materialelor de învățare, discuții cu colegii sau cu un asistent de predare și susținerea evaluărilor sub formă de chestionare formative.

Comportamentele de autoevaluare ale elevilor au fost înregistrate în baza de date cu marcaje temporale. În acest studiu a fost analizată relația dintre comportamentele de autoevaluare ale elevilor și performanța lor de învățare. Scorurile la examenul intermediar pentru curs au fost folosite ca indicator al performanței de învățare. Un total de 73 de elevi au participat la acest experiment. Șapte elevi nu au susținut deloc evaluarea și au fost astfel excluși.

S-a aplicat gruparea ierarhică în clustere pentru a identifica modele specifice de comportament de autoevaluare online. Numărul de cluster a fost determinat cu ajutorul dendrogramei care arată distanța dintre fiecare punct de date. Deoarece fiecare caracteristică are o scară diferită, care poate afecta performanța algoritmilor de grupare, toate caracteristicile au fost standardizate pentru analiza grupării. În continuare, diferența dintre performanța de învățare a elevilor din tiparele identificate a fost examinată folosind un model statistic.

Deoarece distribuția fiecărei caracteristici nu a îndeplinit cerințele pentru efectuarea ANOVA, s-a folosit testul Kruskal-Wallis H, pentru a examina dacă performanța fiecărei caracteristici a diferit semnificativ între grupuri. Descoperirile au arătat diferențe semnificative de comportament în cadrul fiecărui grup, ceea ce sugerează că este fezabilă utilizarea grupării ierarhice pentru a clasifica comportamentul autoevaluat al elevilor. Pentru a înțelege modul în care caracteristicile individuale afectează performanța de învățare, s-a efectuat o analiză de corelație Spearman pentru a analiza relația dintre fiecare comportament și scorul de examinare.

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. Rezultatele 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 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.

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:

Limitarea principală a studiului este dimensiunea mică a eșantionului (76 de studenți), ce a redus generalizarea rezultatelor. Sunt necesare rezultate similare din experimente pe scară largă pentru a confirma relația dintre comportamentele de autoevaluare online ale elevilor și performanța de învățare.

[6] Fizic, Universitate columbiana,

AdaBoost, Bayesian GLM, Decision Trees, Logit Boost, Random Forest și Stochastic Gradient Boosting

Kolmogorov Smirnoff, Random Forest

Scop:

Obiectivul acestui articol este de a prezenta rezultatele implementării unui sistem informațional predictiv (IS) pentru prevenirea abandonului universitar într-o instituție de învățământ superior. Sistemul trebuie să permită calcularea riscului de abandon per elev și să utilizeze o procedură de generare a alertelor pentru coordonarea intervențiilor.

Abordare:

Proiectarea metodologică a constat din 3 faze: colectare de date, modelare și implementare/validare. IS-ul rezultat a fost implementat într-o instituție de învățământ superior cu peste 15.000 de studenți, oferă o interfață cu utilizatorul pentru monitorizarea principalilor factori de risc asociați abandonului școlar, în timp util și personalizat.

Datele au fost colectate, consolidate și completate cu informații obținute de la Ministerul Columbian al Educației Naționale și de la Direcția Columbiană de Dezvoltare Socială.

Datele au constat inițial din 44.031 de înregistrări a 15.805 studenți, acoperind patru perioade academice (2016-1, 2016-2, 2017-1 și 2017-2), grupate în 165 de variabile inițiale diferite.

O analiză a corelației a fost efectuată folosind testul Kolmogorov Smirnoff (KS) pentru fiecare dintre variabilele legate de ratele de abandon. Ulterior, a fost utilizată o metodă de analiză multivariată folosind un algoritm Random Forest pentru a determina gradul de contribuție a diferitelor seturi de variabile în predicția ratei abandonului școlar. Au fost generați arbori de decizie pentru cele mai semnificative variabile. Arborii de decizie au furnizat intervalele sau valorile în care rata abandonului este semnificativ mai mică sau mai mare decât media.

Toate variabilele au fost ordonate în funcție de potențialul lor predictiv. Acest proces necesită utilizarea tehnicilor avansate de Data Mining deoarece, pentru fiecare variabilă, trebuie luată în considerare dependența ei condiționată de prezența altor variabile în ceea ce privește abandonul, și nu numai dependența ei necondițională. Un algoritm Random Forest a fost folosit ca metodă de regresie a datelor.

În loc de a predefini un singur model care funcționează pentru majoritatea instituțiilor de învățământ, au fost testate simultan mai multe metode și, în final, a fost aleasă metoda care se potrivește cel mai bine realității instituției. Astfel, modelul predictiv va fi adaptat în mod specific modelelor de abandon observate la studenții instituției. Algoritmii luați în considerare în această fază au inclus: AdaBoost, Bayesian GLM, Decision Trees, Logit Boost, Random Forest și Stochastic Gradient Boosting.

Aria de sub curba ROC oferă o măsură a performanței modelului pentru variabilele selectate, echivalentă cu probabilitatea ca modelul să clasifice un student real care abandonează cu o probabilitate mai mare decât un student fals care abandonează (fals pozitiv). Cei doi algoritmi cu cea mai mare zonă sub curba ROC (aproape de 80%) au fost algoritmul AdaBoost și GML Bayesian și au oferit cel mai bun echilibru între nivelul predictiv și stabilitatea modelului.

Concluzii:

IS-ul implementat reprezintă un instrument puternic pentru prezicerea, monitorizarea și gestionarea factorilor de risc asociați cu factorii de abandon școlar. Printre avantajele IS-ului implementat se numără centralizarea informațiilor și prioritizarea în urmărirea la timp a studenților cu un risc mai mare de abandon. De asemenea, permite înregistrarea individuală și masivă a intervențiilor efectuate elevilor.

Probleme ramase nerezolvate:

Printre oportunitățile de cercetare viitoare se numără și încorporarea unor noi metode de clasificare și ponderare, care permit îmbunătățirea fiabilității predicțiilor sistemului informațional.

[7] Online, Universitate Israel

Decision Tree, Logistic Regression

CHAID, Mann-Whitney U

Scop:

Studiul își propune să caracterizeze cursanții în funcție de tiparele lor de învățare și să identifice indicatori care prezic succesul elevilor într-un mediu online. Se prezintă un studiu asupra a trei cursuri online de licență generală de chimie oferite la Open University of Israel (OUI).

Finalizarea cursurilor online este cunoscută a fi mai dificilă decât cursurile tradiționale față în față. Studiul are două obiective principale:

1. să caracterizeze cursanții în funcție de modelele lor de învățare în mediul de învățare online,
2. să identifice modele de învățare care pot prezice finalizarea cu succes de către studenți a cursurilor de chimie online.

Deoarece datele cursurilor online prezintă în general informații despre comportamentul de învățare, acest studiu include măsuri care se referă la frecvența deschiderilor de lecții online și a trimiterii temelor, care indică gradul de implicare a cursanților.

Abordare:

Studiul s-a concentrat pe rolul unui factor central care afectează succesul în cursurile online: învățarea autoreglată și implicarea cursanților. În acest scop, s-a folosit o abordare cu metode mixte care combină interviurile semi-structurate și analiza statistică.

Designul cercetării a inclus instrumente atât calitative, cât și cantitative (metode mixte). Studiul a fost realizat în două etape: prima etapă a fost descriptivă și a doua a fost predictivă. Scopul primei etape a fost identificarea comportamentelor de învățare pe care studenții le aplică în cadrul cursurilor online de chimie generală. Această etapă s-a bazat pe interviuri cu participanții înscriși la fiecare dintre aceste cursuri și pe o analiză descriptivă a fișierelor jurnal ale cursurilor. Înainte de a decide ce variabile vor fi folosite în analiza cantitativă, s-a efectuat o fază de preprocesare, colectare, clasificare și filtrare a datelor. În urma acestei etape a început etapa predictivă, în cadrul căreia s-a evaluat cantitativ puterea colectivă a variabilelor majore care au fost identificate în prima etapă ca predictori de succes.

Baza de date a enumerat numeroase activități instantanee ale studenților individuali la anumite cursuri. Pentru a analiza aceste date într-un mod semnificativ, s-au definit două variabile colective, enumerate mai jos, care descriu modelele de învățare ale elevilor:

1. Student Cumulative Opening Pattern (SCOP): SCOP reprezintă sesiuni video cumulative pe care fiecare student le-a deschis până la o săptămână dată, fără a ține cont de deschiderea repetată de către același student (variabilă de interval discret).
2. Starea de încarcare a primei teme opționale: Aceasta este o variabilă binară care face distincția între studenții care au încărcat prima temă opțională și cei care nu au depus-o.

S-au folosit testele Chi-Square și Mann-Whitney U pentru a examina asocierea dintre parametrii măsurați și rata de succes a elevilor.

S-au aplicat două modele de regresie logistică și un algoritm Decision Tree. S-au creat două modele. Pentru modelul A, s-a folosit tehnica Decision Tree, ce este o tehnologie de extragere a datelor care efectuează diverse sarcini de clasificare și regresie și este, în general, considerată eficientă și simplă. Aici s-a folosit Chi-Square Automatic Interaction Detection (CHAID) ca măsură de selecție a atributelor.

Existența modelului B a pornit de la ideea că, pentru a prezice dacă un student va reuși la un curs, este nevoie de o metodologie statistică care ar putea explica un rezultat dihotomic (reușit/nereușit) bazat pe o colecție de categoriale, ordinale și intervale. Abordarea regresiei logistice oferă o astfel de analiză, trecând de la prezicerea apariției unui eveniment la prezicerea probabilității acestuia să se producă.

Modele de regresie logistică au fost construite pe baza datelor colectate de la toate cele trei cursuri. Fiecare model s-a bazat pe datele studenților înscriși în anii 2016–2019 (n = 797). Datele din primul semestru al anului 2020 (n = 157) au fost folosite pentru a valida aceste modele. Dintre cei 797 de studenți, 478 (60%) au depus prima temă opțională. Un test Chi-Square a găsit o asociere semnificativă din punct de vedere statistic între starea de depunere a primei sarcini opționale și succesul general al cursului. Modelul clasifică corect 70% din cazuri.

În plus, utilizând testele Hosmer-Lemeshow Goodness-of-Fit, s-a constatat că regresiile sunt evaluate ca fiind bine adaptate. Aceste rezultate sugerează că atunci când studenții își trimit prima temă opțională, se poate determina probabilitatea ca un anumit student să finalizeze cursul. Modelul clasifică corect 66% din cazuri.

Concluzii:

S-a efectuat o analiză DT separată pentru fiecare săptămână a cursului și s-au actualizat valorile SCOP și rata de depunere a temelor în consecință. Variabilele independente sugerate au fost SCOP-ul fiecărei săptămâni, statutul de depunere a sarcinilor opționale (începând din săptămâna 5), sexu și districtul de reședință. Variabilele independente au fost selectate pentru fiecare săptămână conform algoritmului CHAID.

S-a constatat că acuratețea algoritmului este de 76% utilizând validarea încrucișată de zece ori.

Concluziile acestei cercetări sunt deosebit de relevante în contextul pandemiei Covid-19 și vor continua să fie importante în lumea post-Covid-19. Învățarea online a crescut considerabil în ultimii ani și a devenit extrem de importantă odată cu tranziția rapidă la învățarea online atât în școli, cât și în universități, în urma epidemiei de Coronavirus.

Probleme ramase nerezolvate:

O limitare a acestui studiu provine din considerente etice care au restricționat autorii să conecteze activitățile online ale studenților la datele obținute din interviurile lor. Astfel, inerviurile nu au putut să se relaționeze cu succes la un model de învățare online în cadrul cursului și au rămas folosite doar în cadrul analizei statistice.

În cele din urmă, o limitare importantă este legată de interpretarea unei deschideri video repetate. Mulți studenți au revizionat videoclipurile cursului, dar modelele lor de revizionare nu au putut fi evaluate cu exactitate. Acest lucru se datorează faptului că datele despre deschiderea mai multor videoclipuri au inclus atât studenții care au revizionat lecțiile, cât și pe cei care pur și simplu le-au redeschis din cauza unor probleme tehnice.

[8] Online Universitate, Nitra Slovacia

Random Forest (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosting (GB), Feed-forward Neural Networks (FNN)

Scop:

În această lucrare, autorii propun un nou model predictiv bazat pe un hibrid de Random Forest (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosting (GB) și Feed-forward Neural Networks (FNN) pentru a prezice abandonul studenților din cadrul unei universități.

Abordare:

Setul de date brute utilizat pentru realizarea experimentelor în acest studiu a fost colectat direct de la Universitatea Constantine the Philosopher din Nitra din 2016 până în 2020. Datele inițiale conțin 261 de eșantioane și 12 caracteristici ale studenților înscriși la lecția pregătitoare despre sisteme de baze de date. Variabila de ieșire are două valori, fie 1 pentru non-abandon, fie 0 pentru studenții care au abandonat cursul universitar. Dintre toți elevii, 210 (80,46%) au promovat cursul cu succes și nu au abandonat școala și 51 (19,54) nu au promovat.

Procesul de selecție a caracteristicilor a fost efectuat pentru a decide variabilele de intrare pentru construirea modelului. O hartă de corelație a fost concepută pentru a analiza corelațiile dintre caracteristicile de intrare și variabilele de ieșire. Există o corelație medie între caracteristicile importante și notele studenților. Ca rezultat, caracteristicile puternic corelate sunt luate în considerare în construirea modelului datorită impactului lor cel mai mare asupra rezultatelor studenților.

În acest studiu, se propune un nou model format dintr-un hibrid de RF, XGBoost, GB și FNN pentru a prezice abandonul studenților la orele universitare. Se propune un hibrid de patru modele pentru a face un meta-învățator puternic. Abordarea propusă constă din două straturi. În primul strat, previziunile temporale ale RF, XGBoost și GB sunt generate folosind un set de date de antrenare complet pentru a extrage meritele fiecărui clasificator de bază. În al doilea strat, predicțiile generate în primul strat sunt transmise modelului FNN pentru a calcula predicția finală a abandonului elevilor folosind validarea încrucișată.

Deoarece caracteristicile sunt extrase folosind transformări neliniare complexe în al doilea strat al modelului nostru, clasificatori de complexitate ridicată în stratul de ieșire nu sunt necesari.

Concluzii:

Pe setul de date colectat din 2016 până în 2020 la Universitatea Constantin the Philosopher din Nitra, metoda propusă a demonstrat o performanță mai mare în comparație cu modelele de bază în aceleași condiții. Pe baza constatărilor acestui studiu, elevii cu risc de abandon școlar pot fi identificați pe baza unor factori influenți și diferiți agenți ai educației pot face referire la aceste informații pentru intervenția timpurie și reducerea riscului abandonului școlar.

Modelele de clasificare binară selectate pot fi utilizate pentru a prezice abandonul sau non-abandonul studenților la nivel de curs individual, chiar și atunci când setul de date este limitat și are un număr limitat de caracteristici de intrare.

Probleme ramase nerezolvate:

O limitare a acestui studiu este că în ani diferiți, cursurile, deși au rămas aceleași, au furnizat date diferite pentru a fi analizate. Ca urmare, a fost dificil să se determine care atribute sunt suficient de importante pentru a prezice performanța studenților în general.

Pentru cercetări viitoare, alte abordări, cum ar fi învățarea profundă și alte modele hibride, pot fi utilizate pentru a prezice abandonul studenților și pentru a compara rezultatele cu constatările acestui studiu. Trebuie luați în considerare și alți factori influenți care nu sunt prezentați în acest studiu și se recomandă un studiu suplimentar de analiză a caracteristicilor pentru a ajuta agenții educației să gestioneze cu succes problemele abandonului studenților.

[9] Universitate, Fizic

Retele neuronale, Decision Tree

Correlation Feature Selection, Chi-Square, Gain Information, Random Forest

Scop:

Acest document prezintă un studiu experimental pentru obținerea unui model predictiv care să permită anticiparea abandonului universitar.

Scopul acestui studiu s-a constituit din găsirea unui model bazat pe algoritmi de învățare computațională (arbori de decizie și rețele neuronale) care anticipează dezertarea universitară în scopul reducerii ratei de abandon în programele de licență în Inginerie, Științe Sociale și Administrative.

Abordare:

Studiul folosește 51.497 de instanțe cu 26 de atribute obținute despre studenții înrolați în programe de licență de Științe Sociale, Științe Administrative și Inginerie colectate din 2010 până în 2019.

Acest studiu experimental s-a bazat pe metodologii educaționale de extragere a datelor și algoritmi de învățare computațională, cum ar fi rețele neuronale și arbori de decizie. De asemenea, algoritmi de selecție a atributelor și metode de reeșantionare au fost utilizate pentru a echilibra clasa principală.

Selectarea atributelor importante folosind algoritmi permite îmbunătățirea calității datelor de intrare cu eliminarea atributelor care nu sunt relevante. În acest studiu, se folosesc selecția subsecțiunilor cu CFS (Correlation Feature Selection), metoda filtrului (Chi-Square și Gain Information) și o metodă wrapper (Random Forest). În primul rând, s-au folosit metodele de rang pentru a determina cele mai relevante atribute. În al doilea rând, a fost efectuată o analiză cu metode de filtrare și de wrapper pentru a determina un subset cu cele mai semnificative atribute.

Concluzii:

Rezultatele arată că cel mai performant model a fost cel Random Forest cu un coeficient de corelație Matthew de 87,43% (față de 53,39% obținut de rețelele neuronale artificiale) și o precizie echilibrată de 94,34%.

Pe de altă parte, testele efectuate indică faptul că o posibilă creștere a numărului de neuroni din stratul ascuns al ANN ar putea îmbunătăți performanța alogoritmului RNA. Cu toate acestea, necesită mai multă putere de procesare.

Modelul a permis estimarea unui număr de posibile abandonuri pe perioadă, contribuind la instanțele implicate în prevenirea sau reducerea abandonului în învățământul superior.

Probleme ramase nerezolvate:

Pentru a continua cercetarea, autorii iau în considerare utilizarea unor atribute suplimentare, cum ar fi istoricul plăților, datoriile acumulate, accesul la campus și alte atribute similare.

De asemenea, se intenționează a se folosi alți algoritmi de clasificare și o combinație de clasificatori pentru a genera o soluție mai robustă.

De asemenea, se intenționează a se folosi servicii cloud, cum ar fi Machine Learning în AWS, Azure Machine Learning, BigML, IBM Watson, TensorFlow pentru a îmbunătăți timpul de procesare.

[10] Fizic, liceu portughez

CNN

BPSO

Scop:

Această lucrare propune predicția performanței elevilor folosind metoda de selecție a caracteristicilor bazată pe CNN (Convolution Neural Network) și BPSO (Binary Particle Swarm Optimization). În acest studiu, sunt realizate clasificatoare pentru predicții cu 2 și 5 clase de ieșire. În acest context, acest studiu își propune să prezică performanța elevilor.

Abordare:

Este utilizat un set de date din depozitul de învățare automată UCI. Cu acesta, a fost posibilă compararea tehnicilor deja consolidate în domeniul EDM, cu tehnica de învățare automată.

Extragerea caracteristicilor este un pas important în clasificare, deoarece eficacitatea unui model de învățare depinde de variabilele de intrare (trăsături substanțiale) care descriu caracteristicile elevilor și pot fi utilizate pentru a prezice performanța elevului. Aceste date se referă la performanța elevilor de liceu din două școli portugheze. Datele despre atribute (inclusiv notele elevilor, caracteristicile demografice, sociale și academice) au fost colectate folosind rapoarte școlare și chestionare.

Tehnica BPSO este utilizată pentru a selecta un subset de M caracteristici dintr-un set de N caracteristici ale unei baze de date, (unde M<N), pentru a reduce redundanța în baza de date, astfel încât să poată fi obținute rezultate optime.

În ceea ce privește predicția de performanță, tehnica EDM propusă a fost adecvată, iar care rezultatele obținute sunt următoarele:

* CNN cu o acuratețe de 91,14% în 2 clase pe setul de date matematice,
* CNN-BPSO cu o acuratețe de 93,33% în 2 clase pe setul de date matematice,
* CNN cu o acuratețe de 84,40% în 2 clase pe setul de date matematice,
* CNN-BPSO cu o acuratețe de 86,21% în 2 clase pe setul de date matematice,
* CNN cu o acuratețe de 93,85% în 2 clase pe setul de date portugheze,
* CNN-BPSO cu o acuratețe de 96,6% în 2 clase pe setul de date portugheze,
* CNN cu o acuratețe de 85,58% în 5 clase pe setul de date portugheze,
* CNN-BPSO cu o acuratețe de 86,11% în 5 clase pe setul de date portugheze.

Concluzii:

A fost generat un model pentru algoritmul BPSO-CNN utilizat în acest studiu. Pentru aplicarea algoritmilor au fost utilizate în baza de date toate atributele și caracteristicile selectate BPSO pentru 2 clase de ieșire. Pentru 5 clase (notele A, B, C, D și F) de ieșire. Metodele propuse oferă o precizie mai mare de 93,3 % pentru 2 clase, în timp ce predicția pentru 5 clase este de 86,21 %.

Pentru a compara rezultatele obținute cu metodele propuse în această lucrare cu rezultatele obținute în alte lucrări, s-au efectuat teste cu metodele de clasificare utilizate în unele lucrări. Metoda propusă BPSO-CNN CNN tradițional cu o îmbunătățire de 1,6 % a preciziei. Metoda uzuală bazată pe regresia logistică oferă acuratețe de 62,05% și, respectiv, 67,69% în setul de date la matematică și, respectiv, portugheză, ceea ce este foarte scăzut în comparație cu metoda propusă.

Mai mult, a fost posibilă identificarea faptului că atributele care se referă la notele și absențele elevilor sunt mai predictive pentru performanță decât înregistrările demografice ale elevilor.

Sistemul propus susține o acuratețe superioară de 96,6% față de diferite lucrări de cercetare anterioare și a constatat că majoritatea atributelor importante sunt legate de activitățile școlare în comparație cu datele despre caracteristicile demografice și socioeconomice.

Probleme ramase nerezolvate:

Nu au fost identificate.

[11] Hibrid, Universitate, Beijing

Decision tree - Gini Index (DT-GI), Decision tree - Information Gain (DT-IG), Decision tree - Gini Ratio (DT-GR), Random Forest (RF), Deep Learning (DL), Logistic Regression (LR), Naïve Bayes (NB).

Ranker, Greedy Stepwise

Scop:

Principalele caracteristici cheie ale acestui studiu examinează statutul studenților (promovați sau nu), utilizând notele pe parcursul anului de studiu și notele examenului final pentru determinarea performanțelor de studiu.

În plus, s-a conceput selecția atributelor care se bazează atât pe un filtru, cât și pe o metodă bazată pe wrapper. Principala contribuție a acestui studiu este explorarea celor mai bune caracteristici de impact care joacă un rol vital în prezicerea performanței elevilor

Abordare:

Setul de date este colectat de la sistemul educațional online Smart Learning Partner (SLP) asociat unui centru de inovare avansată pentru educația viitoare, de la Universitatea din Beijing. Analiza a fost realizată cu setul de date care oferă una dintre caracteristicile cunoscute sub numele de jurnalele de învățare. Aceste jurnale de învățare permit studenților să se conecteze la sistem înainte de examenele finale și să efectueze activități precum întrebări, clarificări și interpretare.

Perioada de progres a fost analizată din 2016 până în 2018, iar, în fiecare an, studenții au fost conectați la sesiunea de jurnal de învățare o dată înainte de examenele finale. Datele colectate provin de la 11814 elevi din 31 de școli locale din Beijing.

Variabila țintă este cunoscută sub denumirea de „situație”, care descrie dacă un elev trece sau nu în anul următor de studiu.

Selecția atributelor (AS) este implementată folosind metode bazate pe filtre și wrappere. Metodele bazate pe filtre includ metode de căutare, și anume Ranker și Greedy Stepwise. Metodele de wrapper folosesc combinații ale fiecărei perechi de atribute și evaluează cele mai bune caracteristici care realizează o precizie mai mare. Un model de validare încrucișată de 10 ori este realizat în faza de evaluare a modelului pentru a valida rezultatele clasificării. După aceea, clasificarea bazată pe ansamblu este efectuată prin validarea divizată a setului de date (adică împărțirea 70% antrenament și 30% testare). În cele din urmă, comparația și rezultatele clasificării bazate pe meta-ansamblu față de performanțele modelului de referință (prin validare încrucișată) sunt validate.

Doar șapte modele de clasificare au fost selectate în acest experiment, șapte modele ale căror performanțe au fost bune utilizând validarea încrucișată de 10 ori: Decision tree - Gini Index (DT-GI), Decision tree - Information Gain (DT-IG), Decision tree - Gini Ratio (DT-GR), Random Forest (RF), Deep Learning (DL), Logistic Regression (LR), Naïve Bayes (NB). Modelul Decision tree - Information Gain a obținut cel mai mare scor.

Concluzii:

Se prezintă o nouă abordare de clasificare care realizează selecția caracteristicilor printr-un filtru improvizat și o metodă bazată pe wrapper. Metodele sunt testate pe un set mare de date compus din caracteristici noi în diferite domenii. Cadrul propus validează rezultatele atât prin validare încrucișată cu 10, cât și prin validare divizată. În plus, modelul bazat pe meta-ansamblu identifică performanțele elevilor, indiferent dacă aceștia promovează sau nu la examenele finale.

Probleme ramase nerezolvate:

Nu au fost identificate.