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, Performanta academica

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.

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 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;

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 Promovare-Eșec, Rezultate Excepționale-Eșsec, Rezultate Excepționale-Promovare și, respectiv, Retragere-Promovare.

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

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] 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.

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.

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 studneții 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.

Feature Selection:

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.

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.

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] Studenti, 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:

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.

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%.

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] 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 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.

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

Feature Selection:

Metoda Feature Subset Selection Wrapper a selectat cele mai relevante atribute și a crescut ratele de precizie. Selectarea atriutelor 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.

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] Online

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 într-un curs de contabilitate. Studenții au fost instruiți să finalizeze autoevaluarea online sub formă de chestionare după oră. S-a efectuat o analiză de grupare, care a evidențiat trei modele de comportament în autoevaluarea online și s-a 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 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. 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.

Student self-assessment behaviors were recorded in the database with timestamps. 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. A total of 73 students participated in this experiment. Seven students did not take the assessment at all and were thus excluded.

Feature Selection:

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 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.

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] 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.

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.

Feature Selection:

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.

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 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, Decision Tree

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 și (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. S-au aplicat două modele de regresie logistică și un algoritm Decision Tree și s-au găsit doi parametri care pot prezice finalizarea cursului: starea îndeplinirii unei sarcini opționale și modelul de deschidere video cumulat al elevilor (students’ cumulative video opening pattern - SCOP).

The research design included both qualitative and quantitative tools (mixed methods). 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 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. 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.

The database 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:

1. 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).
2. 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. Therefore, we used Chi-Square and Mann-Whitney U tests to examine the association between the measured parameters and students’ success rate.

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 provides such an analysis by moving from predicting an event occurrence to predicting its probability to occur.

A decision tree (DT) is a data mining technology that carries out various classification and regression tasks and is generally considered effective and simple. Here we used the chi-square Automatic Interaction Detection (CHAID) method as an attribute selection measure, based on the statistical chi-square test for independence.

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. 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. The model correctly classifies 70% of the cases.

Furthermore, using the Hosmer-Lemeshow Goodness-of-Fit tests, 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. The model correctly classifies 66% of the cases, and similar to the model A above, the Hosmer-Lemeshow Goodness-ofFit 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.

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.

The algorithm’s accuracy was found to be 76% using tenfold cross-validation.

Concluzii:

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. 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.

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.

Probleme ramase nerezolvate:

A limitation of this study stems from ethical considerations that restricted the authors from connecting the online activities of individual students to their interview data. Thus, they 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. Many students re-watched the course’s videos, but it their re-watch patterns could not be accurately assessed. This was because data on multiple video opening included both students who re-watched lessons and those that simply re-opened them due to technical issues.

[8] Online

Scop:

In this paper, the authors 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:

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. The initial data contains 261 samples and 12 features of students registered in the preparatory lesson on database systems. The output variable has two values either 1 for non-dropout or 0 for students who 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.

Besides, the feature selection process was conducted to decide the input variables for model building. A correlation map was designed to analyze the correlations between input features and output variables. There exists 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 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. 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 layer are fed to the FNN model to compute the final prediction of student dropout using cross-validation. RF, XGBoost, and GB are chosen as the first layer’s base models in this study.

Because features are extracted using complex non-linear transformations in the second layer of our model, complex classifiers in the output layer are unnecessary.

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.

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 information 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 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.

Probleme ramase nerezolvate:

A 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 student’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 research and trends in LA and EDM provide 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.

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.

[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 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.

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.

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. 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

Concluzii:

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.

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 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 5-class predictions. In this context, this study aims to predict student performance.

Abordare:

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.

Feature extraction 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.

Binary particle swarm optimization 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.

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, 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 %.

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.

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%.

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:

The dataset is collected from the online educational system Smart Learning Partner (SLP) at advanced innovation center for future education, beijing normal university. 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.

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. the data gathering consists of 11814 students from 31 local schools in beijing that were categorized into seven subjects.

The target variable (class label) is known as 'situation', which describes whether a student passes or fails.

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.

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).

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 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. at the same time, remaining attributes' scores 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).

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. 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. 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: