Educational Data Mining: Student Performance Prediction

E2 57/2024 Ivan Partalo E2 19/2024 Milica Vujić

1. Definicija problema

Ideja ovog projekta je da se obrade podaci o faktorima iz studentskog života koji utiču na uspeh studenata. Potrebno je utvrditi koji su ključni faktori koji utiču na uspeh, a potom ih iskoristiti za obučavanje modela koji bi trebao da prepozna obrasce u podacima i predvidi da li student pripada grupi uspešnih, srednje uspešnih ili manje uspešnih studenata, odnosno klasifikuje ih. S obzirom da su podaci o studentima iz Portugala, Pakistana, Iraka i Južno Afričke Republike ideja je i da se analiziraju sličnosti i razlike između faktora koji utiču na uspeh studenata iz različitih država.

2. Motivacija problema rešavanog u projektu

Analizom faktora bi se moglo uočiti koji su faktori presudni za uspeh studenata. Ti podaci bi bili korisni univerzitetima na kojima su podaci prikupljani kako bi unapredili svoju nastavu, videli ko su grupe studenata koje treba dodatno motivisati i kako ih motivisati. Pozitivni trendovi iz jedne zemlje mogli bi biti uzor univerzitetima drugih zemalja kako da poboljšaju studentski život i samim tim svoj generalni uspeh.

3. Relevantna literatura

3.1 Educational Data Mining: Student Performance Prediction in Academic (2019)

Tema rada

Tema je klasifikacija studenata po uspešnosti na studijama u zavisnosti od različitih podataka o njima (radne navike, način života…). U radu su prikazani rezultati različith algoritama za klasifikaciju i njihovo poređenje, kao i kako korišćenjem samo određenih podataka, a izbacivanjem manje bitnih se dolazi do bolje klasifikacije. Cilj je izvući koje stvari najviše utiču na uspešnost studenta na studijama.

Metodologija

Klasifikacioni algoritmi koji su korišćeni su: NaiveBayes, DecisionTree, RandomForest, RandomTree, REPTree, JRip, OneR, SimpleLogistic, ZeroR. Skup podataka sadrži 21 klasu u rezultućoj koloni što je previše klasa za klasifikovanje, pa je to namapirano na 5 klasa.

Podaci

Skup podataka koji se koristi u radu se sastoji od 33 kolone i 649 redova. Kolone obuhvataju razne vrste podataka o studentu: godine, pol, obrazovanje roditelja, učenje na nedeljnom nivou, prisustvovanje na vannastavnim aktivnostima, slobodno vreme posle škole, zdravstveni status itd.

Evaluacija rešenja
Za svaki od upotrebljenih klasifikacionih algoritama metrika koja je korišćena je tačnost. Najveća tačnost ostvarena kada su uključene sve 32 kolone je bila 76.7% za OneR algoritam. Za ostale algoritme tačnost je povećana kada su bile uključene samo 8 najuticajnijih kolona, na primer za Decision Tree algoritam sa 67.8% na 76.6% tačnosti.
Zaključak
U radu se može videti kako je uzimanjem u obzir samo određenih kolona postignuta veća tačnost klasifikacije nego korišćenjem svih kolona, na šta ćemo obratiti pažnju u našem radu.
Mana ovog rada može biti što postoji mali broj redova u skupu podataka (649), ako bi postojalo više redova, mogao bi se napraviti bolji klasifikacioni model. Osim toga nisu korišćene neuronske mreže, što nam daje motiv da ih koristimo u našem radu, kako bi videli da li možemo postići bolji rezultat.

3.2 Student Performance Prediction Model based on Supervised Machine Learning Algorithms (2020)
Tema rada
Glavni cilj je razviti modele za predviđanje: statusa studenata – da li će student položiti ili pasti završni ispit. U radu se vrši i predikcija konačnih ocena studenata – klasifikacija u jednu od šest kategorija (F, P, M, G, V, E), gde je F najlošija, a E najbolja ocena. Istraživanje se fokusira na primenu mašinskog učenja kako bi se unapred identifikovali studenti sa slabim performansama i sprovele se odgovarajuće mere.
Podaci
Podaci dolaze iz kurseva bachelor programa na College of Computer Science and Information Technology, Univerzitet Basra, za akademske godine 2017–2018 i 2018–2019. Nakon preprocesiranja, ukupno je analizirano 499 studenata. Atributi o studentima su godina studija, pol, godina rođenja, da li je obnavljao godinu, da li je zaposlen, bodovi za aktivnost, bodovi na ispitu.
Metodologije
U radu se koristi više algoritama i porede se njihovi rezultati. Decision Tree se koristi za razumevanje ključnih faktora koji utiču na uspeh studenata. Naïve Bayes i K-Nearest Neighbour (KNN) se koriste za klasifikaciju da li je student položio ili ne. Multi-layer Perceptron (MLP) se koristi za uočavanje nelinearnih relacija između faktora i ciljne klase. Support Vector Machine je korišćen za predikciju ocene i klase sa jasnom granicom. Takođe su korišćeni i PART, JRip, OneR, ali su pokazali slabije rezultate.
Evaluacija rešenja
Korišćene su metrike True Positive Rate (TPR), False Positive Rate (FPR), Precision i osetlivost (Recall). Najtačnija procena da li je student položio ili pao je postignuta korišcenjem KNN-a 88,6%, BayesNet algoritma 88.4%, a potom J48 Decision Tree 87.6%.

Zaključak
Na projektu planiramo da koristimo više algoritama, na primer KNN, neki od Decision Tree algoritama, neuronske mreže koji su u ovom radu koji obrađuje sličan problem našem pokazali dobre rezultate.

3.3 Supervised data mining approach for predicting student performance (2019)
Tema rada
Cilj ovog istraživačkog rada je razvoj prediktivnih modela korišćenjem algoritama za klasifikaciju kako bi se predvidela studentska uspešnost, odnosno kategorizacija studenata na odlične i neodlične, u zavisnosti od rezultata njihovih akademskih postignuća.
Podaci
Korišćeni su podaci o studentima osnovnih studija sa Faculty of Computer and Mathematical Sciences at Universiti Teknologi MARA Cawangan Kelantan i Universiti Teknologi MARA Cawangan Negeri Sembilan, prikupljeni tokom 6 semestara u periodu 2013-2016 i uzet u obzir uspeh na 10 ključnih predmeta, kako bi se predvideo konačan uspeh. Ukupno je uzeto u obzir 631 diplomirani studenat i za svakog je poznato ime, student ID, pol, ukupni prosek (CGPA). GCPA se pokazao kao najvažniji atribut za klasifikaciju.
Metodologije
Za konačan uspeh korišćene su dve kateorije: odličan i nije odličan. Skup podataka je predprocesiran kako bi se osiguralo da nema nedostajućih i neželjenih podataka.
Za klasifikaciju su korišćeni algoritmi KNN, Decision Tree, Naive Bayes i Logistic Regression Model. Utvrđeno je da se najveća tačnost sa KNN algoritmom postiže kada se uzme za k 9 kateogorija (A+, A, A-, B+, B, B-, C+, C, C- D+, D, D-, E, F). Od dva korišćena Decision Tree algoritma GINI i Information Gain, Information Gain se ispostavio bolji i proizveo je stablo od 19 čvorova i 16 listova.
Evaluacija rešenja
Za evaluaciju je korišćena mera tačnosti, preciznosti i odziva. Najveća tačnost je dostignuta korišćenjem Naive Bayes algorima 89.26%, a praćen je Logistic Regression algoritmom 85.28%, KNN algoritmom sa 84,8%, i Decision Tree algoritmima.
Zaključak
Ovaj istraživački rad obrađuje sličnu temu koju mi obrađujemo. Značajan je za nas jer se u njemu koriste algoritmi od kojih bismo neke i mi koristili, a i mera tačnosti koju ćemo takođe koristiti. Rezultati rada su značajni za univerzitete da uvide koji faktori utiču na uspeh studenata.

4. Skup podataka

Podaci koje bismo koristili su 4 data set-a iz 4 različite države. Data set-ove smo preuzeli sa interneta i to:

- Podaci o studentima sa Stellenbosch University, South Africa
- Podaci o studentima sa Govt Islamia Gratuate college Kasur, Punjab, Pakistan
- Podaci o studentima iz tri škole iz Iraka
- Podaci o studentima iz Porta, Portugal, prikupljeni za rad studije P. Cortez and A. Silva for the study "Using Data Mining to Predict High School Student Performance"

Podaci su tabelarni i obuhvataju:

- generalne podatke o studentima- pol, godine;
- podatke o porodici - da li učestvuje u školovanju, nivo obrazovanja, veličina porodice;
- koliko prosečno sati posećuje nastavu, koliko prosečno sati uči;
- koliko prosečno sati spava;
- podatke o prihodima studenta;
- da li student ima partnera;
- kakav ima društveni život
- smeštaju (uslovima za učenje)
- uspeh

i ostale podatke koji postoje u nekim data set-ovima, a u drugima ne, poput nivoa stresa, konzumacije alkohola, zanimanja roditelja… Probali bismo da spojimo sva 4 data set-a u jedan veliki, kako bismo imali više podataka za obučavanje modela, gde bismo imali dodatni atribut država, a neke podatke koje imamo samo u nekim državama ne bismo razmatrali.

5. Metodologija

Na početku je potrebno izvršiti pretprocesiranje podataka, u našem slučaju to obuhvata nalaženje zajedničkih kolona u svim skupovima podataka i onda spajanje skupova podataka po tim kolonama.

Studente bismo klasifikovali u kategorije veoma (A), srednje (B) i slabije uspešnih (C), jer je takvo rezultujuće obeležje u skupu podataka o studentima iz Pakistana, koji je najveći od 4 data set-a. U skupovim podataka gde je uspeh izražen u brojčanim vrednostima od 0-100 morali bismo definisati granice kategorija.

Za klasifikaciju studenata po njihovoj uspešnosti koristićemo više klasifikacionih algoritama kako bi videli koji najviše odgovara za rešavanje ovog problema. Za postavljanje baseline-a koristićemo Random Forest algoritam. Nakon toga koristićemo k-NN i neuronske mreže.

6. Metod evaluacije

S obzirom da je broj studenata u sve tri klase okvirno jednak (skup podataka je izbalansiran) koristićemo tačnost kao metriku za evaluaciju naših modela.