Предикција академског успеха студената

Предлог пројекта из предмета Системи за истраживање и анализу података

Дефиниција проблема

Предикција академског успеха студената на основу демографских, социоекономских и макроекономских података, академских података о упису, као и академском успеху на крају првог и другог семестра.

Након експлоративне анализе података, креирања и тренирања модела биће извршена евалуација сваког модела одговарајућим метрикама, како би се упоредили резултати добијени различитим приступима.

Мотивација

Висок проценат академски образованих појединаца је од кључног значаја за економски и социјални развој друштва. Највећи проблем који високошколске установе морају да реше како би унапредиле свој успех јесте проблем одустајања студената од студија. Иако се стопа дипломирања увелико разликује у различитим земљама и институцијама, уопштено говорећи, просечно сваки трећи студент одустаје од уписаних студија. Чак и у Данској, најуспешнијој земљи по питању академског образовања, само око 80% студената заврши студије, док је у Италији ова стопа 46%.

Главни циљ овог рада је да кроз предикцију исхода студирања утврди који фактори у највећој мери доводе до напуштања студија и на тај начин помогне да се идентификују студенти који на основу свог демографског и социоекономског статуса, као и постигнутог успеха током студирања припадају ризичној групи.

Релевантна литература

[1] *Nagy, M., & Molontay, R. (2018, June). Predicting dropout in higher education based on secondary school performance. In 2018 IEEE 22nd international conference on intelligent engineering systems (INES) (pp. 000389-000394). IEEE.*

Задатак рада је предикција исхода студија студената уписаних на 1. годину Универзитета технологије и економије у Будимпешти. Предикција је вршена на основу постигнућа студената у средњој школи и неких личних података (пол, године и слично). Подаци над којима је вршено истраживање су добијени са поменутог Универзитета и обухватали су 15 825 студената уписаних у периоду од 2010. до 2017. године.

Проблем који је требало решити је бинарна класификација студената, при чему циљно обележје има вредност дипломирао или одустао. За проблем класификације, аутори су користили следеће методе машинског учења: стабло одлучивања, Random Forest, Gradient Boosting, Наивног Бајеса, KNN, логистичку регресију, генерализовани линеарни

модел и дубоко учење. Како би смањили димензионалност полазног скупа података, за сваки од наведених алгоритама кориштен је другачији подскуп полазног скупа података. Подскупови су изведени на основу релевантних атрибута за одређене моделе машинског учења. За поређење модела, извршена је 10-струка унакрсна валидација, а кориштене метрике су *accuracy*, *recall*, *precision* и AUC. Најбољи резултати постигнути су моделом дубоког учења, са тачношћу од 73.5%, док је најнижа тачност од 63% постигнута стаблом одлучивања.

Иако је скуп података другачији у односу на скуп који ћемо ми користити, с обзиром на то да не узима у обзир податке о академском успеху у току студија, овај рад нам је релевантан јер се бави истим проблемом који и ми решавамо. При томе, аутори су за решење одабрали велики скуп алгоритама машинског учења и на основу њихових резултата се можемо лакше одлучити за алгоритме које ћемо ми користити у нашем раду.

[2] *Berens, J., Schneider, K., Görtz, S., Oster, S., & Burghoff, J. (2018). Early detection of students at risk–predicting student dropouts using administrative student data and machine learning methods. Available at SSRN 3275433.*

Задатак рада је предикција исхода студија на основу различитих података о студентима применом метода машинског учења. Предикциони модел је обучен и тестиран на подацима прикупљеним са 2 универзитета у савезној држави North Rhine-Westphalia, и то са државног универзитета - State University (SU) и приватног универзитета примењених наука - Private university of applied sciences (PUAS). Што се тиче скупа података добијеног са државног универзитета, он броји податке о 14496 студената, док скуп података са приватног факултета садржи 7600 инстанци. Инстанце су описане атрибутима које можемо груписати на демографске (пол, године, националност и сл.) и академске (везане за успех на студијама – просечна оцена и слично). За предикцију циљног обележја са вредностима 0 (позитиван исход) и 1 (негативан исход), аутори су користили регресиону анализу, неуронске мреже, стабло одлучивања и AdaBoost алгоритам. За евалуацију модела кориштене су метрике *accuracy* и *recall.* Аутори су евалуацију модела за оба скупа података вршили 5 пута – на основу података о студентима приликом уписа и на основу података на крају сваког од 4 семестра. Као алгоритам са највећом тачношћу показао се AdaBoost алгоритам. Користећи само демографске податке доступне у време уписа, предикциони модел достигао је тачност од 67% за SU, док се тачност предикције повећала на 80% у четвртом семестру. Одговарајући резултати за PUAS су само 50% у тренутку уписа и 83% у четвртом семестру. Такође, аутори су закључили да чак бројни демографски подаци не утичу значајно на тачност модела једном када академски подаци постану доступни.

Овај рад нам је од значаја јер се бави истом тематиком као и наш рад. Осим тога, за разлику од претходног рада који посматра искључиво податке пре уписа студија, овај рад користи податке који се односе на успех на студијама, које садржи и наш скуп података. Такође, закључак аутора о пресудном утицају академских у односу на демографске податке ће нам бити од значаја приликом одабира улазних атрибута за наш предикциони модел.

[3] *Pojon, M. (2017). Using machine learning to predict student performance (Master's thesis).*

Задатак рада је примена алгоритама машинског учења у предикцији успешности студената. Акценат рада је на поређењу различитих метода машинског учења, али и *feature engineering*-у који побољшава предиктивну моћ модела. Аутор је користио три методе машинског учења: линеарну регресију, стабло одлучивања и Наивног Бајеса. Такође, кориштена су 2 различита скупа података. Први скуп података садржи податке о 480 студената из разних земаља, углавном са Блиског истока. Подаци садрже 17 атрибута, који су већином прикупљени у току студија (подаци о присуству, као и активном учешћу студената на предавањима, али и разним дискусијама), али такође садржи и неке основне социоекономске податке. Други скуп података садржи 395 инстанце описане 31 атрибутом, који су за разлику од првог скупа, већином социоекономске природе (степен образовања и занимање родитеља, као и аспекти друштвеног живота студента – учесталост излазака, конзумирање алкохола током радне седмице и викенда и слично). Изабрани алгоритми су примењени на оба скупа података и то 2 пута – на сирове податке и након примене *feature engineering*-а. Као најбољи алгоритам показао се Наивни Бајес за 1. скуп података, са тачношћу од 97.5% и стабло одлучивања за други скуп података, са тачношћу од 77.6%, након примене *feature engineering*-а, док је тачност наведених алгоритама у случају необрађених података 95.8% и 68.4%, респективно.

Осим што се бави истом тематиком, овај рад нам је значајан због начина евалуације, као и због примене *feature engineering*-а за побољшање перформанси предикционог модела. Најважнији закључак из овог рада јесте да је *feature engineering* важнији фактор у постизању боље предикције модела у односу на избор алгоритма, када су у питању подаци коришћеним у овом раду, што можемо прилагодити и на наш скуп података. Такође, слично као и у претходном раду, показано је да се много већа тачност модела постиже када се он обучава на подацима прикупљеним у току студија, у односу на социоекономске податке.

Скуп података

Скуп података који ћемо користити доступан је на Kaggle-u, на линку: https://www.kaggle.com/datasets/ankanhore545/dropout-or-academic-success .

Подаци се односе на евиденцију студената уписаних од академске 2008/2009. године (након примене Болоњског процеса на високо образовање у Европи) до 2018/2019. Укључују податке са 17 основних студија из различитих области, као што су агрономија, дизајн, образовање, медицина, новинарство, менаџмент, техничке науке. Подаци се могу груписати у неколико категорија: демографски, социоекономски, макроекономски, академски подаци приликом уписа, академски подаци на крају првог и другог семестра.

Детаљан преглед свих категорија, атрибута који им припадају, њихових вредности и расподеле доступан је на линку: https://valoriza.ipportalegre.pt/piaes/features-info-stats.html .

Скуп података садржи 4424 инстанце описане са 37 атрибута, при чему 36 атрибута представљају улазе у модел, а атрибут *target* је категорички атрибут, са вредностима Graduate, Dropout i Other чија је дистрибуција 50%, 32% i 18%, респективно.

Методологија

Прва фаза у изради пројекта биће експлоративна анализа података. Први корак биће визуализација података (исцртавање хистограма дистрибуције атрибута) и рачунање основних статистичких вредности за податке (средња вредност, медијана, дисперзија, минимум, максимум). Први проблем који се уочава на основу визуализације података јесте неравномерна расподела циљног атрибута, који има вредност Graduate у случају 50% инстанци. Овај проблем можемо решити већ на нивоу података, употребом неке од техника узорковања попут Synthetic Minority Over Sampling Technique (SMOTE) или Adaptive Synthetic Sampling Approach (ADASYN). Такође, у оквиру експлоративне анализе, ради бољег разумевања података са којим радимо, извршили бисмо кластеровање. Као алгоритам за кластеровање бисмо користили KNN, а број кластера бисмо одредили помоћу *Elbow* методе.

Због великог броја атрибута у скупу података, биће неопходно извршити *feature engineering*, како бисмо утврдили који атрибути не утичу значајно на предиктивну моћ модела, а како бисмо смањили његову комплексност. За ту сврху можемо користити *Yellowbrick* библиотеку и њене визуализаторе података попут *Rank Features*, који рангира појединачне или парове атрибута за детекцију коваријанси. Рангирање може бити 1D или 2D. Више детаља о поменутој библиотеци, може се наћи на линку https://www.kaggle.com/code/parulpandey/analysing-machine-learning-models-with-yellowbrick/notebook . Након припреме података, приступићемо решењу проблема, применом модела машинског учења. Наш проблем се своди на проблем класификације, а као класификационе моделе користићемо логистичку регресију, стабло одлучивања, Random Forest, AdaBoost, Gradient Boosting, Bagging, Наивног Бајеса и вештачку неуронску мрежу.

Када добијемо коначан модел, извршићемо његову интерпретацију коришћењем ELI5 библиотеке и PDP-а (PD, *Partial Dependence Plot*). ELI5 библиотека омогућава да утврдимо која обележја су најважнија за перформансе модела, а на основу 1D и 2D PDP графика можемо установити како промена појединачног обележја утиче на вредност циљног обележја.

Метод евалуације

Скуп података ћемо поделити на тренинг, валидациони и тест скуп у односу 70:20:10. Модел ће бити трениран над тренинг скупом, а оптимизација хипер-параметара ће бити извршена над валидационим скупом, применом *grid search*-а. За евалуацију модела биће коришћене метрике из релевантне литературе попут тачности (*accuracy*), прецизности (*precision*), одзива (*recall)* и F1 мере. Поменуте метрике биће израчунате за сваки модел и за сваку комбинацију хипер-параметара и након тога ће бити извршено њихово поређење. При томе, као главна метрика за евалуацију и оптимизацију хипер-параметара узимаће се F1 мера. Када утврдимо најбољу комбинацију вредности на основу тренинг и валидационог скупа, формираћемо коначан модел који ћемо применити над тест скупом, а затим извршити његову евалуацију. Резултате које добијемо ћемо визуализовати матрицом конфузије, чиме ћемо почети анализу грешака модела. На основу матрице конфузије моћи ћемо да уочимо обрасце у грешкама - да ли је већина погрешно класификованих инстанци класификована у доминанту класу и слично.

План

Израда пројекта би требало да обухвати следеће фазе (*milestones*):

- Експлоративна анализа података,
- Креирање модела,
- Верификација модела и прилагођавање параметара,
- Анализа добијених резултата.

Тим

- Марија Кљештан (E2 56/2022)
- Лука Матић (E2 109/2022)