Metodološki radovi
UDK 37.022/.026:005.334

ПРИМЈЕНА DATA MININGA ЗА ПРЕДВИЂАЊЕ УСПЈЕХА У СТУДИРАЊУ

APPLICATION OF DATA MINING FOR PREDICTING SUCCESS IN LEARNING

Владо Симеуновић

Резиме: Рад се бави креирањем модела за предвиђање успјешности студената у току студирања помоћу DATA MININGA, те анализом фактора који утичу на остварени ниво успјешности. Креиран је модел који на основу социо-демографских података о студентима, те података о њиховом понашању и ставовима према учењу и организацији наставе у цјелини, настоји класификовати студенте у једну од двије категорије успјешности. Успјешност је мјерена оствареном просјечном оцјеном студената у периоду студирања. Тестирали смо двије методе data mininga i to: логистичку регресију и стабла одлучивања. Сматрали смо да је ће представљени модел послужити као тест за стварње шире базе ажурираних података уз коришћење неког од информатичких алата, те да ће се на основу њега дефинисати велики број атрибута којима ће се релативно поуздано предвиђати успјешност у студирању

Кључне ријечи: успјех у студирању, стабла одлучивања, логистичка регресија, data mining

Abstract: *The work deals with creating of model to predict the performance of students during the study using data mining, and analysis of factors affecting the achieved level of performance. The model which is created on the bases of socio-demographic data on students, and data on their behavior and attitudes towards learning and teaching process organization as a whole, seeks to classify students in one of two categories of performance. Performance is measured by students average grade achieved in the period of study. We tested two methods of data mining as follows: logistic regression and decision trees. We considered that the presented model would be served as a test for creation of broader base of updated data by usage some of the information tools, and that on the basis of this model will be defined a number of attributes that would relatively reliable predict the study preformance*

Key words: *success in the study, decision trees, logistic regression, data mining*

1. УВОД

Предвиђање (прогноза) будућности је круна сваке науке. Образовање има стратешку важност за економски и друштвени развој, тј. за развијање друштва заснованог на знању. У процесу европских интеграција неопходно је образовани систем ускладити са критеријумима и препорукама Европске уније или других европских организација и процеса, уз поклањање посебне пажње индикаторима успјешности образовног система које је ЕУ дефинисала. За високообразовне установе анализа успјешности студирања врло је важна, јер стратегијско планирање студијских програма претпоставља

проширивање или смањење обима или дубине изучаваних садржаја, као и мијењање структуре васпитно-образовног процеса у зависности од успјешности студената. Успјешност студирања на факултетима до сада је углавном истраживана у циљу проналажења просјечних оцјена, дужине студирања и сличних показатеља, док фактори који утичу на постизање успјеха нису довољно истражени. Постоје развијени модели за предвиђање успјешности који могу помоћи при одлуци о прихватању кандидата за упис на студије, а који углавном укључују демографске податке о студентима, иако се наглашава значај укључивања и других информација о апликантима. [10]

Циљ овог рада је проналажење важних чинилаца који утичу на успјех студената, који је представљен просјечном оцјеном. У ову сврху смо употријебили двије методе рударења података које су погодне за класификацију: логистичку регресију и стабла одлучивања. Логистичка регресија је статистички поступак заснован на расподјели вјероватноће која се показала ефикасном у многим подручјима предикције. Њихова тачност упоређена је са стаблима одлучивања како би се идентификовао модел који даје тачнију класификацију студената. Резултати истраживања се заснивају на анкетном истраживању спроведеним са студентима Педагошког факултета у Бијељини академске 2009/2010. године, при чему су, осим социо-демографских података о студентима, прикупљени и подаци о њиховом успјеху, али и ставови о организацији рада на факултету. На основу тих података креиран је каузални модел са демографским и другим карактеристикама студената као улазним варијаблама, те просјечном оцјеном у претходној академској години као излазном варијаблом.

Анализа значајности варијабли добијених логистичком регресијом и стаблом одлучивања указују на јачину утицаја поједине улазне варијабле на успјех студената, на основу чега је могуће донијети закључак о могућим предикторима успјешности студирања.

Рад се састоји од приказа употријебљене методологије, прегледа претходних истраживања у том подручју, те од приказа резултата и закључка са смјерницама за будућа истраживања.

2. ПРИМЈЕНА СТАБАЛА ОДЛУЧИВАЊА У ПРОЦЕСУ ПРЕДВИЂАЊА

У циљу изградње што успјешнијег модела, на посматраном узорку тестирана је једна од непараметријских метода рударења података: стабло одлучивања, тачније њихова подврста – класификациона и регресиона стабла (eng. Cassification And Regression Trees – CART). Овом методом добија се графички приказ модела утицаја улазних варијабли на излазну, која је

изражена у облику класа или категорија. Сваки чвор у графичком стаблу представља једну улазну варијаблу, на чијим су рубовима означена „дјеца-чворови“ за сваку могућу вриједност неке улазне варијабле. Сваки лист у стаблу представља вриједност циљне (излазне) варијабле, ако су дате вриједности улазних варијабли представљене путем од коријена, стабла до тог листа. Стабло се добија „учењем“ на подацима, на начин да се врши гранање (eng. splitting) изворног скупа података у подскупове на основу тестирања вриједности варијабли. Процес се понавља на сваком изведеном подскупу на рекурзивни начин (eng. recursive partitioning). Рекурзија је завршена када подскуп одређеног чвора има све исте вриједности излазне варијабле, или када даље гранање више не придоноси побољшању резултата. [9]

За изградњу стабла коришћен је CART алгоритам, према Breiman et al. у [9], који на основу расположивих података о улазним и излазним варијаблама креира бинарно стабло гранањем слогова у сваком чвору, а према функцији одређеној за сваку улазну варијаблу. Евалуациона функција коришћена за прелом је Гини индекс (ИГ), дефинисан према формули [1]:

$$I_G(t) = 1 - \sum_{i=1}^{m} p_i^2$$

(једначина 1)

гдје је m тренутни чвор, p_i је вјероватноћа класе i у чвору t, а m је број класа у моделу (у нашем случају m=2).

Алгоритам CART узима у обзир сва могућа гранања како би пронашао оно најбоље за тачност модела. Најбоље гранање одређује се за сваки атрибут у сваком чвору, а побједник се бира помоћу Гини индекса. Алгоритам може успјешно да ради са континуираним и категоријалним варијаблама.

Стабло расте све док се не пронађе ново гранање које побољшава његову успјешност у раздвајању слогова у класе. Будући да свако сљедеће гранање има на располагању мање репрезентативну популацију, потребно је смањивати стабло (eng. pruning), како би се добила тачнија класификација. Циљ је идентификовати оне гране које омогућују најмање предиктивне способности по листу у грани, како бисмо их избацили из стабла. У процедури смањивања стабла скупови грана смањивани су у односу на комплетно почетно стабло одлучивања, што је процедура слична елиминисању предиктора у дискриминативној анализи. На крају је изабрано стабло одговарајуће величине с обзиром на тачност класификације. При томе се узима у обзир однос сложености стабла и величине грешке. Побједничко подстабло се одабира на основу укупне грешке (стопе погрешне класификације) добијене када се модел примијени на тестираном узорку.

Стабло у овом раду креирано је на основу четрнаест улазних категоријалних варијабли.

3. ПРИМЈЕНА ЛОГИСТИЧКЕ РЕГРЕСИЈЕ У ПРОЦЕСУ ПРЕДВИЂАЊА

Логистичка регресија или логистички модел, односно, логит модел се користи за предвиђање вјероватноће догађаја путем прилагођавања података логистичкој кривој. Логистичка регресија је тип регресионе анализе у којој је зависна (критеријумска) промјењива дихотомна, односно бинарна и кодира се са 0 или 1 и постоји најмање једна независна (предикторска) промјењива.

3.1.Тумачење модела логистичке регресије

Статистичко моделовање бинарних промјењивих подразумијева мјерење избора које за сваки субјекат може бити успјешно или неуспјешно. Бинарни подаци су вјероватно најчешћи облик категоријских података. Најраспрострањенији модел бинарних података је *логистичка регресија*.

За бинарни избор Y и квантитативну објашњавајућу промјењиву X, нека $\pi(x)$ представља вјероватноћу успјеха када X има вриједност x. Ова вјероватноћа је параметар за биномну дистрибуцију. Модел логистичке регресије има линеарни облик за логит ове вјероватноће.

$$\log it[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$$

(једначина 2)

Ова формула приказује да $\pi(x)$ расте или опада са С-функцијом од *x*.

Друга формула за логистичку регресију односи се директно на вјероватноћу успјеха. Ова формула користи експоненцијалну функцију $\exp(x) = e^x$ у облику

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

(Једначина 3)

3.2.Тумачење линеарне апроксимације

Параметар β одређује стопу раста или опадања С-криве. Ознака β указује на то да ли је крива опадајућа или растућа, као и на стопу раста промјене како $|\beta|$ расте. Када модел има вриједност $\beta = 0$, десна страна једначине 3 поједностављује се у константу. Затим, $\pi(x)$ је идентичан са свим *x*, те крива

прелази у хоризонталну праву линију. Бинарни избор Y постаје потом константа X.

3.3. Тумачење рација вјероватноће дешавања

Наредно тумачење модела логистичке регресије користи вјероватноћу дешавања и рација вјероватноће дешавања. Као модел вјероватноће избора (тј, изгледи за успјех) користити се сљедећа једначина:

$$\frac{\pi(x)}{1-\pi(x)} = \exp(\alpha + \beta x) = e^{\alpha}\left(e^{\beta}\right)^{x}$$

(једначина 4)

Експоненцијални однос пружа тумачење за β: изгледи се повећавају мултипликативно за e^{β} за свако повећање од једне јединице по *x*. Другим ријечима, вјероватноћа на нивоу x+1 једнака је вјероватноћи при *x* помножено са e^{β}. Када је β = 0, $e^{\beta} = 1$, тада се вјероватноћа не мијења како се мијења вриједност *x*.

Логаритам вјероватноће, што представља логит трансформацију π(x), има линеарни однос. Овде се ради о логит изразу модела, што говори да се логит повећава уз β јединицу за сваку јединицу промјене при *x*. Већина не схвата логит скалу као нешто природно, тако да она има ограничену употребу.

3.4.Тест значаја

Код модела логистичке регресије, нулта хипотеза X_0 : β = 0 значи да је вјероватноћа успјеха независна од X. Код већих узорака, статистика теста

$$z = \frac{\beta'}{ASE}$$

има стандардну, нормалну дистрибуцију када је β = 0. Уз то, *з* се може придодати стандардној табели да бисмо добили једностарну или двострану П-вриједност. Исто тако, за двострану алтернативу β ≠ 0, $(\beta' / ACE)^2$ важи Валдова статистика код које важи кси-квадратна дистрибуција великог узорка са *∂ф = 1*.

Иако Валдов тест добро функционише код великих узорака, тест рација вјеродо- стојности је ефектнији и поузданији за величине узорка које користимо у пракси. Статистика теста пореди максимални $Л_0$ лог-функције вјеродостојности када је β = 0 (то јест, када π(x) мора да буде идентична са свим вриједностима *x*) до максималног $Л_1$ лог-функције вјеродостојности за

нерестриктивну β. Статистика теста, -2(Л0 – Л1), такође, има кси-квадратну дистрибуцију великог узорка са *дф = 1*. Већина софтвера за логистичку регресију даје податке за максималну лог-вјеродостојност $Л_0$ и $Л_1$, а статистика рација вјеродостојности добија се из ових максима.

3.5. Дистрибуција прорачуна вјероватноће

Процијењена вјероватноћа да је Y = 1, при фиксном скупу x од X износи

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

(једначина 5)

Већина софтвера за логистичку регресију може да прикаже процјене као и интервале поузданости за праве вјероватноће. Интервали поузданости за вјероватноћу се могу извести коришћењем матрице коваријансе модела процјене параметра. Услов α’ + β’x у експонентима једначине предвиђања (Једначина 3) је процијењени линеани предиктор у логит трансформацији πx. Процијењени логит има велики узорак АСЕ дат процијењеним коријеном на квадрат од

$$Var(\alpha' + \beta' x) = Var(\alpha') + x^2 Var(\beta') + 2xCov(\alpha', \beta').$$

Интервал поузданости од 95 посто за прави логит је (α’ + β’x) ± 1,96АСЕ. Замјеном крајњих тачака интервала за α + βx у експонентима Једначине 4 добија се одговарајући интервал вјероватноће. Може се догодити да се занемари модел уклапања, те да се једноставно користе узорци пропорција да би се процијениле такве вјероватноће.

Када модел логистичке регресије функционише, естиматор вјероватноће заснован на моделу далеко је бољи показатељ од узорка пропорције. Овај модел има само два параметра за процјену, гдје приступ који није заснован на моделу има одвојени параметар за сваку дистинктивну вредност X. Стварност је мало компликованија. У пракси, неће сваки модел тачно представљати стварни однос између π(x) и x. На примјер, како се повећава величина узорка, естиматор заснован на моделу не мора да се приближава у потпуности тачној вриједности вјероватноће. Међутим, то нужно не мора да значи да је узорак пропорције заиста бољи естиматор у пракси. Ако се модел приближава стварној вјероватноћи на одговарајући начин, његов естиматор и даље тежи да буде ближи правој вриједности од пропорције узорка. Овај модел филтрира податке узорка. Резултирајући естиматори теже да буду бољи осим ако свака пропорција узорка није заснована на екстремно великом узорку. И коначно, ако модел логичке регресије приближно одговара правој зависности π(x) на x, тада су тачка и прорачун интервала за π(x) засновани на њему прилично корисни.

4. ПРЕГЛЕД ПРЕТХОДНИХ ИСТРАЖИВАЊА

Истраживања у подручју употребе интелигентних метода за предвиђање успјешности студената углавном су оријентисана на развој модела који ће се користити као помоћ при одлучивању о пријему студената на студије. [9] , [6] Такви модели као критерије узимају у обзир информације о кандидату које су расположиве прије уписа, као нпр. завршена средња школа, успјех у средњој школи, социјални статус и друге информације прије студија, те уз помоћ статистичких метода или метода вјештачке интелигенције настоје пронаћи модел који ће продуковати што већу тачност у предвиђању. Иако се неуронске мреже учестало десетинама година успјешно користе у бројним подручјима, посебно након појаве вишеслојне перцептрон мреже, за коју је доказано да може апроксимирати било коју континуирану функцију, у образовању су интелигентне методе више усмјерене на развој туторских система, а мање на развој предиктивних модела успјешности студирања. [5] Једни од првих аутора у том подручју који користе неуронске мреже су Хардгрев (Hardgrave и др.) [10] који упоређују неуронске мреже и традиционалне статистичке технике у предвиђању успјешности студената на дипломском студију.

Даљи рад истих аутора наглашава како се одлука о томе да ли примити неког студента на студије заснива на бројним факторима, те да је нужно развити предиктивне моделе који ће омогућити неком факултету уписивање оних студената за које постоји висока вјероватноћа да ће студирати успјешно. [10] Вилсон и Хардгрев (Wilson и Hardgrave) [10] показују да регресиона анализа није довољно квалитетна у предвиђању успјеха или неуспјеха студента, па поред регресије тестирају и методе класификације, попут дискриминативне анализе, логистичке регресије и неуронских мрежа.

Њихово истраживање показује сљедеће: (а) класификационе технике су погодније за предвиђање успјеха студената од предиктивних метода; (б) предвиђање успјеха или неуспјеха студената на дипломском студију није довољно тачно ако се користе само типични подаци који описују студента и (в) непараметријске процедуре, као што су неуронске мреже, продукују барем једнако тачан резултат као и традиционалне методе и вриједан су потенцијал за даља истраживања у том подручју.

Истим проблемом одлучивања о пријему кандидата на студије бавили су се и Наик и Реготман (Naik и Ragothaman) [6], који су истраживали успјешност на МБА студију. Користили су неуронске мреже, логит и пробит моделе за предвиђање успјешности студената који се упишу на МБА студије. Неуронске мреже су класификовале студенте у успјешне и неуспјешне на основу њиховог просјека оцјена на преддипломском студију, резултатима ГМАТ теста, смјера на преддипломском студију, старости и других

варијабли. Резултати показују да су неуронске мреже једнако успјешне као и остале технике, али их због бројних предности препоручују за употребу у том подручју.

Истраживање Салимена и Мохезара (Sulaimana и Mohezara) [16] бави се истом тематиком, али иде корак даље у идентификовању кључних фактора успјешности. Њихов модел показао је да је досадашњи просјек оцјена студента најзначајнији предиктор његове даље успјешности, док варијабле попут старости, етничке припадности, пола, те година радног искуства нису значајне за успјешност студирања. Шарлаф (Shulruf) и др. [14] проучавали су корелације између добијених показатеља од стране New Zealand National Certificate of Educational Achievement (NCEA), стандардних квалификација студената, те просјека оцјена постигнутог на првој години студија на једном великом универзитету на Новом Зеланду. Након тога су упоредили резултате добијене за Нови Зеланд са онима добијенима на Cambridge International Examinations (CIE), затим са међународним системом стандарда који је усвојен на Новом Зеланду, те са универзитетским просјецима оцјена студената који су уписани на основу тих квалификација. У истраживању такође процјењују алтернативне моделе за доношење одлуке о упису на студије, те њихове импликације на различите групе студената. Најбољи модел који су добили даје предност квалитету и мјеродавности резултата NCEA. Аутори предлажу комбинацију таквог модела и стандардних квалификација која ће осигурати боље резултате на студију.

Зеидах и Далила (Zaidah i Daliela) [8] су упоредили неуронске мреже, стабла одлучивања и линеарну регресиону анализу у предвиђању успјешности студената. Успјех су мјерили кумулативним просјеком оцјена током студија, а као улазне варијабле користили су демографски профил студената и просјек оцјена на првом семестру преддипломског студија. Резултати показују да су све три методе произвеле тачност већу од 80%, док неуронске мреже дају већу тачност од осталих двију метода. Оладокан и сарадници (Oladokun и др.) [12] користили су неуронске мреже за предвиђање успјешности студената на Универзитету Ибадан у Нигерији. Користили су вишеслојну перцептрон мрежу, која је на тестном узорку тачно предвидјела успјех код 74% студената. Као улазне варијабле кориштене су оцјене, комбинација изборних предмета, успјех на тестовима, старост при упису, образовање родитеља, тип и локација завршене средње школе, пол и слично. Излазна варијабла изражена је кроз три категорије успјеха: добар, просјечни, слаб. Релевантност појединих инпута није анализирана.

Из наведеног прегледа претходних истраживања може се закључити да различите методе рударења података могу послужити за предвиђање успјеха на студијама те се могу користити у зависности од образовног система.

5. МЕТОДОЛОШКИ ОКВИР ИСТРАЖИВАЊА

5.1. Циљеви и задаци истраживања

Циљ овог рада је проналажење важних чинилаца који утичу на успјех студената, који је представљен просјечном оцјеном. За ову сврху смо употријебили двије методе рударења података погодне за класификацију: логистичку регресију и стабла одлучивања, те смо имали намјеру тестирати квалитет сваке од њих.

5.2. Хипотезе и варијабле истраживања

Методе рударења података (data mining) омогућују релативно прецизно предвиђање успјеха студената на Педагошком факултету на основу процјене вјероватноће учешћа појединих варијабли. Варијабле које улазе у модел:

1. Критеријумска варијабла: постигнути успјех: до 7,5 – мање успјешни, од 7,51 до 10.00 – успјешни
2. Независне варијабле:

- пол;
- мјесто студирања;
- подаци о стипендији;
- вријеме посвећено учењу;
- материјали, извори и средства који се користе за учење;
- присуство предавањима;
- присуство вјежбама;
- присуство колоквијумима;
- став о важности оцјене коју ће студент добити на испиту;
- квалитет предавања;
- квалитет вјежби;
- квалитет наставних планова и програма;
- квалитет наставника и
- квалитет процеса вредновања знања.

5.3. Обухват истраживања

- Популација (сви студенти Педагошког факултета у Бијељини).
- Узорак 234 студената друге, треће и четврте године студија.

5.4. Обрада података

- Логистичка регресија
- Стабла одлучивања J-48 и ЦАРТ

6. РЕЗУЛТАТИ ИСТРАЖИВАЊА

6.1. Примјена логистичке регресије у предвиђању успјешности студената у студирању

Постоји неколико метода процјене у логистичкој регресији али најчешћи, можда и најмање ризичан у смислу потврђивања хипотезе, је METHOD=BSTEP(LR), за *Stepwise* анализу уназад. Метод се састоји на могућности тестирања „log-likelihood-a“ (вјероватноће) са датом промјенљивом испуштеном из једначине. У табели 1. приказана је укупна статистика тестираних случајева.

Табела 1: Сумарни приказ обрађених резултата

Непондерисани случајви	N	Постотак
Укупан број случајева	234	100.0
Изгубљени случајеви	0	.0
Укупно	234	100.0
Некласификовани случајеви	0	.0
Укупно	234	100.0

Укупан тест модела дат је у табели „омнибус тестови коефицијената модела“. У нашој логистичкој регресији БСТЕП(ЛР) на почетку све промјенљиве су ушле у једначину, а затим је модел тестиран у десет корака. Као што се види, на почетку су све вриједности дате као „корак“, „модел“ и „блокирање“ једнаке са нивоом значајности ,000. У почетном кораку све варијабле су у моделу. У другом кораку дошло је до елиминације једне варијабле која није статистички значајна (,946). У трећем кораку још једна варијабла је елиминисана из модела са степеном значајности (,859), у четвртом кораку је елиминисана варијабла са степеном значајности (,816), у петом кораку варијабла са степеном значајности (,565), а у шестом кораку варијабла са степеном значајности (,487), у седмом кораку варијабла са степеном значајности (,272) и у осмом кораку варијабла са степеном значајности (,117). Кроз поступак од десет корака хи – квадрат тест се постепено смањивао, што је и императив модела, тако да смо од почетне вриједности 68,789 добили умањену вриједност која износи 64,223.

Табела 2: Омнибус тест модела

		χ^2	df	Sig.
Корак 1	Корак	69.696	14	.000
	Резултат	69.696	14	.000
	Модел	69.696	14	.000
Корак 2	Корак	-.003	1	.956
	Резултат	69.693	13	.000
	Модел	69.693	13	.000
Корак 3	Корак	-.007	1	.935
	Резултат	69.686	12	.000
	Модел	69.686	12	.000
Корак 4	Корак	-.068	1	.795
	Резултат	69.618	11	.000
	Модел	69.618	11	.000
Корак 5	Корак	-.064	1	.801
	Резултат	69.554	10	.000
	Модел	69.554	10	.000
Корак 6	Корак	-.358	1	.550
	Резултат	69.197	9	.000
	Модел	69.197	9	.000
Корак 7	Корак	-.300	1	.584
	Резултат	68.897	8	.000
	Модел	68.897	8	.000
Корак 8	Корак	-1.013	1	.314
	Резултат	67.884	7	.000
	Модел	67.884	7	.000
Корак 9	Корак	-1.205	1	.272
	Резултат	66.679	6	.000
	Модел	66.679	6	.000
Корак 10	Корак	-2.456	1	.117
	Резултат	64.223	5	.000
	Модел	64.223	5	.000

Из претходне табеле нисмо могли закључити које су варијабле биле елеминисане из модела. Тек се увидом у табелу бр. 3 могу уочити правилности у елиминацији појединих варијабли из једначине. Прва варијабла која је елиминисана из модела је „пол“. У сљедећем кораку

елиминисана је варијабла „квалитет вјежби“. У трећем кораку елиминисана је варијабла „присуство предавању“. У четвртом кораку учешће варијабле „квалитет предавања“ није значајно учествовало у побољшању вјероватноће укупног модела. У сљедећим корацима су елиминисане варијабле сљедећим редослиједом: квалитет програма, начин учења и мјесто студирања.

Табела 3: Варијабле које не улазе у једначину

			Резултат	дф	Степен значајности
Корак 2	Варијабле	квалитет вјежби	.003	1	.956
	Укупна статистика		.003	1	.956
Корак 3	Варијабле	пол	.007	1	.935
		квалитет вјежби	.003	1	.955
	Укупна статистика		.010	2	.995
Корак 4	Варијабле	пол	.005	1	.946
		квалитет вјежби	.003	1	.960
		квалитет наставника	.068	1	.795
	Укупна статистика		.077	3	.994
Корак 5	Варијабле	пол	.007	1	.932
		присуство предавању	.064	1	.800
		квалитет вјежби	.001	1	.970
		квалитет наставника	.063	1	.802
	Укупна статистика		.141	4	.998
Корак 6	Варијабле	пол	.003	1	.957
		присуство предавању	.072	1	.788
		квалитет вјежби	.017	1	.895
		квалитет програма	.355	1	.551
		квалитет наставника	.126	1	.723
	Укупна статистика		.497	5	.992
Корак 7	Варијабле	пол	.001	1	.971
		присуство предавању	.070	1	.792
		квалитет предавања	.300	1	.584
		квалитет вјежби	.002	1	.966
		квалитет програма	.170	1	.680
		квалитет наставника	.042	1	.837
	Укупна статистика		.795	6	.992
Корак 8	Варијабле	пол	.001	1	.976
		присуство предавању	.063	1	.802
		квалитет предавања	.094	1	.759
		квалитет вјежби	.043	1	.836
		квалитет програма	.482	1	.488
		квалите тнаставника	.280	1	.597
		оцјењивање	1.008	1	.315
	Укупна статистика		1.806	7	.970

Корак 10	Варијабле	пол	.004	1	.951
		начин учења	1.201	1	.273
		присуство предавању	.067	1	.796
		квалитет предавања	.110	1	.741
		квалитет вјежби	.011	1	.917
		квалитет програма	.309	1	.578
		квалитет наставника	.162	1	.687
		оцјењивање	.698	1	.404
	Укупна статистика		2.989	8	.935
Корак 11	Варијабле	пол	.050	1	.823
		мјесто студирања	2.453	1	.117
		начин учења	.579	1	.447
		присуство предавању	.139	1	.709
		квалитет предавања	.027	1	.869
		квалитет вјежби	.003	1	.955
		квалитет програма	.437	1	.509
		квалитет наставника	.137	1	.711
		оцјењивање	.827	1	.363
	Укупна статистика		5.332	9	.804

У коначном моделу процјене вјероватноће учествују сљедеће варијабле: важност оцјене (0.000), присуствовање колоквијумима (0.003), стипендија (0.039), присуство вјежбама (0.049), дужина учења (0.089).

У наредној табели приказани су псеудо R – kvadrati. Cox & Snell индекс имају вриједност од 0 до ,75, тек се са Nagelkerke индексом врши корекција и доводи ниво у опсег од 0 до 1. Наравно да се овдје R не може посматрати као коефицијент детерминације у линеарној регресији, јер је овдје ријеч о пропорционалном учешћу појединих варијабли у укупној вјероватноћи. Регресијом корак по корак у свакој новој етапи укупан резултат обухваћене варијансе се повећавао. У финалном моделу Cox & Snell износе (0,24), а корекцијом Nagelkerke индексом добија се вриједност (0,324) што се може сматрати задовољавајућим исходом.

Табела 4: Коефицијент детерминације

Корак	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	247.118	.258	.347
2	247.121	.258	.347
3	247.127	.258	.347
4	247.195	.257	.347
5	247.259	.257	.347

Корак	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
6	247.617	.256	.345
7	247.916	.255	.344
8	248.930	.252	.339
9	250.134	.248	.334
10	252.591	.240	.324

У табели бр. 5 приказане су предвиђене вриједности зависних промјењивих базираних на моделу цијеле логистичке регресије. Ова табела показује колико је случајева тачно прогнозирано, а колико није. Циљ регресије у десет корака био је да се повећа проценат успјешног предвиђања. У првом кораку смо имали 28 случајева који су требали имати вриједност 1, а добили су вриједност 2 и 30 случајева који су у укупном броју од 138 случајева који су требали добити вриједност 2, добили вриједност 1. Дакле, укупна варијанса тачног предвиђања износи 74,8%. У коначном моделу, од укупно 96 случаја који су требали имати вриједност 1, само 26 случај добио је вриједност 2, а од 138 случајева који су требали узети вриједност 2, њих 33 је узело вриједност 1.

Табела 5: Процентуална тачност успјешног предвиђања

Посматрано			Предвиђање		
			успјех		Проценат тачности
			1.00	2.00	
Корак 1	успјех	1.00	68	28	70.8
		2.00	30	108	78.3
	Укупна тачност				75.2
Корак 2	успјех	1.00	67	29	69.8
		2.00	30	108	78.3
	Укупна тачност				74.8
Корак 3	успјех	1.00	67	29	69.8
		2.00	30	108	78.3
	Укупна тачност				74.8
Корак 4	успјех	1.00	67	29	69.8
		2.00	30	108	78.3
	Укупна тачност				74.8
Корак 5	успјех	1.00	67	29	69.8
		2.00	28	110	79.7
	Укупна тачност				75.6

Корак 6	успјех	1.00	64	32	66.7
		2.00	26	112	81.2
	Укупна тачност				75.2
Корак 7	успјех	1.00	63	33	65.6
		2.00	24	114	82.6
	Укупна тачност				75.6
Корак 8	успјех	1.00	71	25	74.0
		2.00	33	105	76.1
	Укупна тачност				75.2
Корак 9	успјех	1.00	71	25	74.0
		2.00	34	104	75.4
	Укупна тачност				74.8
Корак 10	успјех	1.00	70	26	72.9
		2.00	33	105	76.1
	Укупна тачност				74.8

Да не бисмо остали на површини квантитативних података, извршили смо детаљнију статистичко-математичку анализу сваког податка. Помоћу наредне једначине израчунали смо вјероватноћу сваке промјењиве која значајно утиче на степен успјешности на студијама:

$$P(x) = \frac{e^{b0+b1*x}}{1 + e^{b0+b1*x}}$$

(једначина 6)

У десетом кораку добили смо вриједност константе b_0 (b_0 = - 3.685). Константа улази у експоненцијалну функцију као први члан. Израчунали смо вјероватноћу сваке варијабле која је ушла у једначину. Највећу вјероватноћу P(x)=0.88 има предиктор – важност оцјене, чији је b коефицијент 0.943; затим слиједи присуство колоквијуму (P(x)=0.85; b_1=0.999); стипендија је по тежини на трећем мјесту у једначини (P(x)=0.70 b_1= 0.488); присуство вјежбама (P(x)=0.61 b_1=0.631) дужина учења (P(x)=0.59 b_1=-0.378).

Уколико погледамо учешће појединих варијабли у укупном нивоу вјероватноће предвиђања, може се констатовати сљедеће: да редовно присуство колоквијумима пет пута повећава вјероватноћу успјеха на студијама (Exp. B 5.280); уколико им је оцјена коју ће добити важна онда се степен успјешности повећава четири пута (Exp. B 4.218); редовно присуство вјежбама три и по пута повећава степен успјешности (Exp. B 3.250); ако примају стипендију онда се успјешност студената повећава два и по пута (Exp. B 2.590) и ако уче бар два сата дневно биће успјешнији једанпут (Exp. B 1.059).

Ради поткрепљења резултата добијених регресионом анализом може се констатовати да 85% успјешних студената редовно присуствује колоквијумима у односу на 53% присуства мање успјешних студената. Успјешни студенти у 88% случајева присуствују вјежбама у односу на 63% студената из друге групе; њих 79% се изјаснило да им је врло важно коју ће оцјену добити у односу на 45% студената из групе мање успјешних; 85% њих учи у просјеку од два до пет сати дневно у односу на 60% студената друге групе; и на крају 27% њих прима стипендију у односу на 11% из друге групе.

Резултати добијени логистичком регресијом у нашем истраживању показују 74,8% тачног предвиђања успјешности студената у студирању на Педагошком факултету. Варијабле које највише доприносе укупној вјероватноћи предвиђања се могу сврстати у двије групе и то: дидактичку (присуство колоквијумима, присуство вјежбама, дужина учења и важност оцјене) и социјална група (примање стипендије).

6.2. Примјена стабла одлучивања у предвиђању успјешности студената у студирању

За истраживање успјешности студената у студирању на Педагошком факултету у Бијељини користили само CART стабло одлучивања које се у неким од претходних истраживања показало као погодно у предвиђању успјеха студената.

```
*******************************************************
* 2. алгоритам CART
*******************************************************
CART Decision Tree
важност оцјене < 1.5
|  колоквијум < 1.5: 1(67.0/35.0)
|  колоквијум >= 1.5: 2 (27.0/9.0)
важност оцјене >= 1.5: 2(76.0/20.0)
Number of Leaf Nodes: 3
Size of the Tree: 5
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances         166               70.9402 %
Incorrectly Classified Instances        68               29.0598 %
Kappa statistic                          0.4014
Mean absolute error                      0.3873
Root mean squared error                  0.4601
Relative absolute error                 79.9955 %
Root relative squared error             93.5313 %
Total Number of Instances              234
```

```
=== Detailed Accuracy By Class ===
TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
 0.656    0.254     0.643       0.656   0.649       0.683     1
 0.746    0.344     0.757       0.746   0.752       0.683     2
=== Confusion Matrix ===
  a   b   <-- classified as
 63  33 |   a = 1
 35 103 |   b = 2
**************************************************
```

Из датог прегледа видљиво је да је просјечна стопа класификације коју даје стабло одлучивања 70,1%, што је мање од просјечне стопе класификације добијене помоћу логистичке регресије (74,8%). Стабло је посебно тачно при препознавању „лошијих“ студената са нижим просјеком оцјена од 7,5, гдје стопа исправне класификације износи 74,6%. Добијена је нешто нижа стопа класификације за класу 1 - „бољих“ студената (65,6%).

Резултат класификације студената од стране стабла одлучивања на подузорку за тестирање може се илустровати и помоћу матрице конфузије, која у колонама приказује стварни број студената који припадају категорији са нижим (2) или вишим (1) просјеком, док је у редовима приказан број студената које је модел стабла одлучивања сврстао у категорију 2 или 1. На дијагонали матрице конфузије могуће је видјети број студената које је модел исправно класификовао. Из табле је видљиво да укупно 35 студената са нижим просјеком (категорија 2), стабло одлучивања не успијева сврстати у исправну категорију, док је 103 успјешно класификовано. Са категоријом 1 ситуација је другачија, 63 студента је правилно класификовано, док је њих 33 сврстано у погрешну категорију.

Цијели модел је правилно сврстао 166 студената од укупно 234. Стабло је нешто прецизније од регресионе анализе и издваја само двије варијабле које утичу на укупну вјероватноћу и то: присуство колоквијумима и важност оцјене.

7. ЗАКЉУЧАК

Оба модела data mininga пружају могућност успјешног предвиђања успјеха студената. У неким ранијим истраживањима, када су коришћени само социјал-демографски подаци о студентима, модел је имао већи ниво тачности од оног кога смо ми добили у нашем истраживању. Резултати које смо добили одсликавају стање нашег образовног система у области високошколског образовања. Потпуно је јасно да је реформа система изазвала промјене понашања код студената. Оно што се може сматрати забрињавајућим је чињеница да великом броју студената није стало коју ће оцјену добити на испиту, да им није важно да ли ће присуствовати предавањима, јер очигледно

немају високо мишљење о могућностима учења у току самог васпитно-образовног процеса.

Показује се да успјешни студенти више пажње посвећују учењу, учествују на колоквијумима, посјећују вјежбе (што говори о апликативности студија) и да им је стало коју ће оцјену добити. Овдје леже могуће интервенције у процесу моделовања рада на факултету. Наравно да ове факторе треба повезати са могућношћу добијања стипендије.

Потребно је извршити анализу свих добијених резултата, систем предвиђања обогатити новим варијаблама и онда извршити корекције у реализацији васпитно-образовног процеса. Неке од препорука за побољшање васпитно образовног процеса које су произашле из овог истраживања би се могле сврстати у сљедеће групе:

- интензивирати наставни процес;
- чешће организовати рад у мањим групама;
- боље повезивање теоријских са практичним садржајима;
- редовно праћење напретка студената кроз колоквије и друге облике провјере знања;
- пружити подршку студентима при прилагођавању на студије;
- обимом и садржајем осавременити литературу која ће бити прилагођена циљевима и начину провјере знања;
- мотивисати наставнике с циљу подизања нивоа посвећености настави и студентима;
- подићи ниво и квалитет знања наставника о методици наставног рада и вредновању успјеха;
- унаприједити комуникационе вјештине и посвећеност наставника наставном раду и
- осавременити систем евалуације наставе и наставника као један од начина праћења квалитета и подршке промјенама.

Како побољшати модел предвиђања успјешности студената на студију?

- увођењем нових варијабли, нпр. исхода учења;
- повећањем узорка;
- у узорак укључити и друге учитељске и педагошке факултете;
- креирањем интелигентног система подршке систему образовања на високошколским установама.

ЛИТЕРАТУРА

[1] Apte C, Weiss S., *Data Mining with Decision Trees and Decision Rules. Future Generation Computer Systems*, Vol. 13, 1997., str.197-210.
[2] Dunham, M., *Data Mining - Introductory and Advanced Topics*, Prentice Hall, 2003.

[3] Glasser, W. *Control Theory*.,Harper and Row. New York, 1984.
[4] Gojkov, G., *Dokimologija*, Beograd: Uciteljski fakultet, 1997.
[5] Masters, T., *Advanced Algorithms for Neural Networks*, A C++ Sourcebook, John Wiley & Sons, 1995.
[6] Naik, B., Ragothaman, S., *Using Neural Networks to Predict MBA Student Success*, College Student Journal, Vol. 38, No. 1, 2004, str.143-150.
[7] Kirckby, R., *WEKA Explorer User Guide for Version 3-3-4*, University of Waikato 2002.
[8] Zaidah, I., Daliela, R., *Predicting students' academic performance, comparing artificial neural network, decision tree and linear regression*, 21st Annual SAS Malaysia Forum, 5th September 2007, Kuala Lumpur, str. 1 – 6.
[9] Witten I.H., Frank E., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation,* Morgan Kaufman Publishers: San Francisco, 2000.
[10] Hardgrave, B.C., Wilson, R.L., Kent, K.A. *Predicting Graduate Student Success: A Comparison of Neural Networks and Traditional Techniques*, Computers & Operations Research, 21, 1994., str. 249 – 263.
[11] Han, J., Kamber, M., *Data Mining – Concepts and Techniques*, Morgan Kaufman Press 2001.
[12] Oladokun, V.O., Adebanjo, A. T., Charles-Owaba, O. E., *Predicting Students' Academic Performance using Artificial Neural Network, A Case Study of an Engineering Course*, The Pacific Journal of Science and Technology, Vol. 9. No. 1., 2008, str. 72 –– 79.
[13] Rodić, N. (2000): *Latentna struktura uspešnosti diplomiranih studenata Učiteljskog fakulteta u Somboru*, Sombor: Norma, VI; 3: 25-44; Beograd: Nastava i vaspitanje, L, 1: 98 – 113.
[14] Shulruf, B., Hattie, J., Tumen, S., *The Predictability of Enrolment and First-Year University Results from Secondary School Performance*, The New Zealand National Certificate of Educational Achievement, Studies in Higher Education, Vol. 33, No. 6, 2008., str. 685 – 698,
[15] Симеуновић, В. (2005) *Информатика, Методологија , Статистика*, Висока школа унутрашњих послова, Бања Лука
[16] Sulaiman, A., Mohezar, S., *Student Success Factors, Identifying Key Predictors*, Journal of Education for Business, Vol. 81, No.6, 2006., str. 328 – 333.
[17] Suzić, N., *PEDAGOGIJA ZA XXI VIJEK* TT centar, Banja Luka, 2005.
[18] Šipka, P., *Zbirka radova sa područja kriterijuma,* Beograd: Odeljenje za psihologiju Vojnomedicinske akademije,1981