Psihološka obzorja / Horizons of Psychology, 17, 4, 51-72 (2008)

Znanstveni empiričnoraziskovalni prispevek

Odkrivanje zakonitosti in podatkovno rudarjenje v psihologiji: uporaba odločitvenih dreves za napovedovanje dosežka na Lestvici iskanja dražljajev

Andrej Kastrin[*]
Univerzitetni klinični center Ljubljana, Inštitut za medicinsko genetiko, Ljubljana

Povzetek: Odkrivanje zakonitosti iz podatkov je interdisciplinarno raziskovalno področje, ki združuje tehnologije in znanja s področij statistike, podatkovnih zbirk, strojnega učenja in umetne inteligentnosti. Najpomembnejši element procesa odkrivanja zakonitosti iz podatkov je podatkovno rudarjenje. Namen prispevka je dvojen. Prvič, strokovno psihološko javnost želimo opozoriti na kvalitativni preskok v znanstvenem raziskovanju, ki se je začel z uveljavitvijo področja odkrivanja zakonitosti iz podatkov, in drugič, na primeru odločitvenih dreves želimo bralcu približati uporabnost metod podatkovnega rudarjenja v psihologiji. Uporabo odločitvenih dreves ilustriramo z gradnjo napovednih modelov dosežka na Zuckermanovi Lestvici iskanja dražljajev (SSS-V) na osnovi medosebnih razlik v bazičnih potezah osebnosti in lastnostih temperamenta. Prediktorske spremenljivke so bile operacionalizirane na osnovi Eysenckovega osebnostnega vprašalnika (EPQ) in slovenske priredbe Strelauovega vprašalnika temperamenta po Pavlovu (SVTP). Ustreznost odločitvenih dreves za napovedovanje dosežka na lestvici SSS-V smo primerjali s klasičnim statističnim modelom multiple linearne regresije. Z vidika napovedne točnosti se je kot najbolj uspešen sicer izkazal multipli regresijski model, kljub temu pa so odločitvena drevesa primerna metoda za začeten pregled podatkov, vizualizacijo in opis podatkovnih zakonitosti z lahko razumljivimi formalizmi.

Ključne besede: odkrivanje zakonitosti iz podatkov, podatkovno rudarjenje, psihološko ocenjevanje

Knowledge discovery and data mining in psychology: Using decision trees to predict the Sensation Seeking Scale score

Andrej Kastrin
University Medical Centre Ljubljana, Institute of Medical Genetics, Ljubljana, Slovenia

Abstract: Knowledge discovery from data is an interdisciplinary research field combining technology and knowledge from domains of statistics, databases, machine learning and artificial intelligence. Data mining is the most important part of knowledge discovery process. The objective of this paper is twofold. The first objective is to point out the qualitative shift in research methodology due to evolving knowledge discovery technology. The second objective is to introduce the technique of decision trees to psychological domain experts. We illustrate the utility of the decision trees on the prediction model of sensation seeking. Prediction of the Zuckerman's Sensation Seeking Scale (SSS-V) score was based on the bundle of Eysenck's personality traits and Pavlovian temperament properties. Predictors were

[*]Naslov / Address: asist. Andrej Kastrin, univ. dipl. psih., Univerzitetni klinični center Ljubljana, Inštitut za medicinsko genetiko, Šlajmerjeva ulica 3, 1000 Ljubljana, Slovenija, e-pošta: andrej.kastrin@guest.arnes.si

operationalized on the basis of Eysenck Personality Questionnaire (EPQ) and Slovenian adaptation of the Pavlovian Temperament Survey (SVTP). The standard statistical technique of multiple regression was used as a baseline method to evaluate the decision trees methodology. The multiple regression model was the most accurate model in terms of predictive accuracy. However, the decision trees could serve as a powerful general method for initial exploratory data analysis, data visualization and knowledge discovery.

Key words: knowledge discovery from data, data mining, psychological assessment

CC = 2240

Psihologija je bila v preteklosti ena od gonilnih znanstvenih disciplin, ki so usmerjale tok razvoja statističnih metod. Psihološki statistiki danes zasedajo pomembne položaje v različnih mednarodnih statističnih združenjih. Po osnovni izobrazbi je psiholog nenazadnje tudi vodilni manager podjetja SPSS in avtor priljubljenega statističnega paketa SYSTAT. Poleg široko sprejete Thurstonove tipologije merskih lestvic je psihologija najmočnejši pečat pustila z uvedbo metode faktorske analize. S paradigmo odkrivanja zakonitosti iz podatkov (angl. *knowledge discovery*) ter hkratnim razvojem hitrih in učinkovitih metod rudarjenja po podatkih (angl. *data mining*) so se ponudile nove možnosti za analizo podatkov, ki jih klasične korelacijske metode ne omogočajo. Pregled uporabe metod podatkovnega rudarjenja po posameznih znanstvenih področjih v svetovnih bibliografskih zbirkah kaže, da je njihova uporaba v teoretičnih in aplikativnih psiholoških raziskavah še razmeroma velika neznanka. Namen prispevka je zato dvojen. Prvič, strokovno psihološko javnost želimo opozoriti na kvalitativni preskok v metodologiji znanstvenega raziskovanja, ki se je začel z uveljavitvijo področja odkrivanja zakonitosti iz podatkov, in drugič, na praktičnem primeru želimo bralcu približati uporabo nekaterih metod podatkovnega rudarjenja v psihologiji.

Kot odgovor na izzive zajemanja, shranjevanja, modeliranja in upravljanja s podatki in znanjem se je v zadnjem desetletju uveljavilo raziskovalno področje, ki se imenuje odkrivanje zakonitosti iz podatkov (Witten in Frank, 2005). Odkrivanje zakonitosti v podatkih je proces odkrivanja vzorcev in modelov, opisanih s pravili in drugimi človeku lahko razumljivimi formalizmi za predstavitev znanja. Gre za odkrivanje implicitnih, doslej neznanih in potencialno uporabnih zakonitosti iz podatkov, z namenom učinkovitejšega odločanja, razvrščanja in napovedovanja. Pod isto streho združuje znanje, tehnologije in metode, razvite na področjih statistike, strojnega učenja, podatkovnih baz, vizualizacije podatkov, razpoznavanja vzorcev in umetne inteligentnosti. Proces odkrivanja zakonitosti v podatkih poteka v več stopnjah: (i) vzorčenje in selekcija podatkov, (ii) transformacije surovih podatkov, (iii) podatkovno rudarjenje in (iv) interpretacija rezultatov in indukcija spoznanj ter splošnih zakonitosti iz podatkov. Računsko najintenzivnejši del tega procesa predstavlja podatkovno rudarjenje, ki vključuje uporabo metod, tehnik in orodij za

avtomatsko konstrukcijo vzorcev, modelov in zakonitosti iz podatkov.

Podatkovno rudarjenje ne nadomešča klasične statistične obdelave podatkov, ampak predstavlja njeno dopolnitev. Statistično testiranje ali, bolje rečeno, preverjanje statističnih domnev je sestavni del slehernega empiričnega znanstveno-raziskovalnega prispevka. Cilj raziskovalcev je, da z uporabo statističnih testov potrdijo (ali ovržejo) postavljene raziskovalne domneve ter s tem podkrepijo svoje teoretične modele. Paradigma statističnega testiranja je prinesla mnoge nove metode, tehnike in algoritme za analizo podatkov. Težko bi našli raziskovalni problem, ki ga ne bi bilo moč ustrezno obdelati s pomočjo statističnih testov. Poplava statističnih testov je po drugi strani močno prispevala tudi k temu, da so se raziskovalci bolj kot s podatki svojih študij začeli ukvarjali s statističnimi testi samimi, z njihovo konstrukcijo, preverjanjem veljavnosti in uporabnosti. Raziskovalci na področju psihologije so prvi začutili ujetost v stroge okvire klasične statistike in predlagali uporabo od velikosti vzorca neodvisnih mer velikosti učinka (Rosnow in Rosenthal, 1989), s katerimi je bilo moč preseči konservativnost klasične statistike pri odločitvenih problemih. Cohen je temelje kritike klasičnega testiranja statističnih domnev predstavil v članku s pomenljivim naslovom »The Earth is round ($p < .05$)« (Cohen, 1994).

Dolgoletni delitvi statistične metodologije na deskriptivno in inferenčno statistiko se je pridružila ortogonalna razdelitev metod na tiste, ki omogočajo eksploratorno oz. raziskovalno analizo, in na metode, namenjene konfirmatorni analizi (Berthold in Hand, 2007). Za razliko od zapletenih postopkov konfirmatorne analize (npr. modelov strukturnih enačb, analize variančno-kovariančnih struktur, analize poti itd.) eksploratorna analiza ponuja zbir orodij za hitro in učinkovito rudarjenje po podatkih, iskanje vzorcev in zakonitosti. Vsem dobro poznani primer metode za eksploratorno analizo podatkov je npr. metoda razvrščanja v skupine. Pojem eksploratorne analize podatkov lahko v veliki meri izenačimo s pojmom podatkovnega rudarjenja (Berthold in Hand, 2007). Metode podatkovnega rudarjenja ponujajo raziskovalcu (i) močno alternativo klasičnemu pristopu testiranja domnev in (ii) zagotavljajo večjo usmerjenost na podatke in zakonitosti, ki se v podatkih skrivajo.

Glavna razlika med klasičnim statističnim pristopom in podatkovnim rudarjenjem je v tem, da klasična statistična obdelava daje največji poudarek preverjanju raziskovalnih domnev s pomočjo statističnih testov, metode podatkovnega rudarjenja pa omogočajo nov vpogled v raziskovalne probleme (tvorjenje hipotez), samodejno indukcijo interpretacij, spoznanj in zakonitosti. Klasična statistika sledi hipotetično deduktivnemu modelu znanstvenega raziskovanja, metode podatkovnega rudarjenja pa induktivnemu modelu. Še pred leti je veljalo, da so metode podatkovnega rudarjenja namenjene izključno analizi velikih podatkovnih zbirk. Ta trditev danes ne vzdrži več, saj praksa potrjuje dobro obnašanje metod tudi na majhnih podatkovjih (Kononenko in Kuhar, 2007). Metode podatkovnega rudarjenja so v psihologiji razmeroma slabo poznane. V slovenskem prostoru so bile objavljene le tri študije, ki so se problemov s psihološko vsebino lotile z uporabo metod podatkovnega rudarjenja (Gasar, Bohanec in Rajkovič, 2002; Kobal Grum idr., 2004; Slivar, 2008).

Slika po svetu ni nič boljša. Letos je bila v Kanadi organizirana prva mednarodna konferenca s področja podatkovnega rudarjenja v izobraževanju, na kateri je bilo predstavljenih tudi nekaj prispevkov s področja psihologije in psihometrije (Baker, Barnes in Beck, 2008).

Področje odkrivanja zakonitosti iz podatkov je izrazito interdisciplinarno. V Sloveniji se z njim intenzivno ukvarjata dve raziskovalni skupini: Odsek za tehnologije znanja na Institutu Jožef Stefan ter nekoliko mlajši Laboratorij za umetno inteligenco na Fakulteti za računalništvo in informatiko v Ljubljani. Na tem mestu ne bo odveč zgodovinska opomba, da so bili prav psihologi tisti, ki so skupini na »Štefanu« v njenih rosnih letih pomagali s svojim bogatim statističnim in širšim metodološkim znanjem (M. Kline, osebna komunikacija, 2006). Večina učbenikov s področja podatkovnega rudarjenja ponuja bralcu pregled nad pisano paleto metod in algoritmov, ki jih lahko uporabimo nad različnimi tipi merskih lestvic. Zainteresiran bralec bo odličen pregled metod našel v preglednih monografijah (Kononenko in Kuhar, 2007; Witten in Frank, 2005). V nadaljevanju podrobneje predstavimo metodo odločitvenih dreves, ki se zaradi jasnosti pristopa, enostavnosti in možnosti grafične predstavitve rezultatov zdi najbolj primerna za uporabo na problemskih področjih, ki jih označujemo kot »mehka« in kamor sodi tudi psihologija.

Odločitvena drevesa

Odločitveni problem je problem klasifikacije oz. kategorizacije novih oz. neznanih dražljajev v vnaprej določene razrede oz. kategorije. Odločitvena drevesa spadajo v družino hierarhičnih odločitvenih modelov, ki so bili skozi leta razviti v okviru raziskav na področju psihologije, odločitvene analize, operacijskih raziskav in sorodnih področjih (Bohanec, 2006).

Odločitveno drevo je hierarhično urejena struktura, sestavljena iz vozlišč in listov (Mitchell, 1997). Vozlišča cepijo odločitveno drevo v veje, vsaka veja pa se konča z listom. Vsako vozlišče ustreza eni neodvisni spremenljivki X_i in predstavlja odločitev za nadaljevanje poti po drevesu navzdol. Vozlišča razbijejo celoten prostor vrednosti neodvisnih spremenljivk na zaključene podmnožice (Mitchell, prav tam). Pot do napovedi odvisne spremenljivke Y začnemo v korenskem vozlišču na vrhu drevesa. Vsaka pot v drevesu od korenskega vozlišča do lista ustreza enemu odločitvenemu pravilu. V vsakem vozlišču se odločimo za nadaljevanje poti po naslednji veji v skladu z vrednostjo neodvisne spremenljivke, ki sestavlja odločitveno pravilo d_i v tem vozlišču. Ko pridemo do lista drevesa, napovemo vrednost odvisne spremenljivke Y z ustreznim pravilom $Y = d(X_1, X_2, \ldots, X_p)$. Vsak list drevesa vsebuje eno pravilo d_i za napoved vrednosti odvisne spremenljivke Y, ki velja le za tisto podmnožico prostora vrednosti neodvisnih spremenljivk, ki je definirana z zaporedjem odločitev, ki pripeljejo do tega lista. Gradnja oz. indukcija odločitvenega drevesa vključuje fazo gradnje drevesa, ki ji sledi faza rezanja. Pri gradnji drevesa kot

mero izbora neodvisne spremenljivke v vozlišču *i* izberemo ustrezno hevristiko (npr. funkcijo informacijskega prispevka, statistično pomembnost razlike v podmnožicah primerov, ki ju določata veji vsakega vozlišča), za konec indukcije drevesa pa postavimo kriterij, ko nobena neodvisna spremenljivka ne inducira delitve na osnovi hevristike (Mitchell, prav tam). Rezanje je potrebno zaradi prevelikega prileganja tako zgrajenega drevesa učnim podatkom. Porezano drevo je preglednejše in hkrati bolj točno pri napovedovanju novih primerov. Odločitveno drevo, pri katerem je odvisna spremenljivka na nominalni (opisni) merski ravni, imenujemo klasifikacijsko drevo, v primeru numeričnega tipa spremenljivke pa govorimo bodisi o regresijskem bodisi modelnem drevesu. Pri slednjem vsak list predstavlja linearno regresijsko enačbo, pri regresijskem drevesu pa pričakovano vrednost odvisne spremenljivke za primere iz učne množice podatkov, ki pripadajo področju vzorčnega prostora, ki ga pokriva ta list (Witten in Frank, 2005).

Indukcija klasifikacijskega drevesa zahteva, da je zaloga vrednosti odvisne spremenljivke definirana na nominalni merski ravni. V praksi to pomeni, da moramo zvezno odvisno spremenljivko diskretizirati. Diskretizacija zvezne spremenljivke v diskretno v informacijskem smislu sicer predstavlja določeno izgubo informacije, vsebovane v izvorni zvezni spremenljivki, po drugi strani pa prinaša tudi dve pomembni prednosti: (i) diskretne spremenljivke so bolj primerne za opisovanje slabše strukturiranih in manj formaliziranih podatkov, s kakršnimi se srečujemo tudi v psihologiji; (ii) nad diskretnimi podatki lahko v primerjavi z zveznimi podatki uporabimo večje število statističnih metod, hkrati pa so bolj primerni za vizualizacijo podatkov.

Uporabo odločitvenih dreves bomo ilustrirali z rudarjenjem po podatkovni zbirki rezultatov raziskave, v kateri smo s pomočjo psihometričnega instumentarija merili izraženost poteze iskanja dražljajev, bazičnih potez osebnosti in lastnosti temperamenta.

Konstrukt iskanja dražljajev

Iskanje dražljajev je bazična poteza osebnosti, ki se nanaša na vedenja, katerih skupni imenovalec je iskanje novih, raznolikih, kompleksnih in intenzivnih dražljajev ter izkušenj (Zuckerman, 1994). Poteza je večrazsežni konstrukt in jo lahko razmeroma dobro opišemo s štirimi dimenzijami: (i) iskanjem vznemirjenja in pustolovščin, ki se nanaša na iskanje draženja skozi motorične aktivnosti; (ii) iskanjem doživetij, ki se izraža skozi iskanje novih izkušenj na kognitivnem in emocionalnem nivoju; (iii) dezinhibicijo, ki je operacionalizirana skozi vedenja, ki nakazujejo slabo socialno inhibicijo, ter (iv) nagnjenostjo k dolgočasju, ki se nanaša na nepotrpežljivost v znanih, predvidljivih in ponavljajočih se situacijah ter socialnih odnosih. Začetek znanstvenega proučevanja poteze so sprožili odmevni vojaški eksperimenti s področja senzorne deprivacije v petdesetih in šestdesetih letih prejšnjega stoletja (Zuckerman, 1969). Oba glavna teoretična modela, ki poskušata pojasniti medosebne razlike v

izraženosti poteze, je razvil Marvin Zuckerman (1994, 2005); prvi pojasnjuje razlike s procesi kortikalnega vzdraženja in budnosti ter se močno naslanja na Eysenckovo delo v sedemdesetih letih prejšnjega stoletja, drugi, psihobiološki model osebnosti pa je splošnejši in razlike pojasnjuje s prepletom genetskih, nevroloških, psihofizioloških in biokemičnih dejavnikov. Klinično je potrjena stabilna negativna povezava med izraženostjo poteze in koncentracijo znotrajceličnega encima monoaminooksidaze (Zuckerman, 1994, 2005). Študije monozigotnih dvojčkov sicer nakazujejo na visoko stopnjo dednosti poteze (Stoel, De Geus in Boomsma, 2006), vendar pa genetski mehanizmi še niso popolnoma jasni (Munafò, Yalcin, Willis-Owen in Flint, 2008). Z nomološkega vidika je poteza sestavni del bazične osebnostne dimenzije, ki jo Zuckerman opredeljuje kot psihoticizem – impulzivno nesocializirano iskanje dražljajev (Zuckerman, 1994, 2005). Poleg teoretičnih, v preverjanje konstruktne veljavnosti usmerjenih študij, so se raziskovalci ukvarjali tudi s proučevanjem odnosa poteze do različnih vedenjskih vzorcev: zlorabe drog (Bardo, Donohew in Harrington, 1996) in alkohola (Andrev in Cronin, 1997; Wiesbeck idr., 1996), prehranjevalnih navad (Pliner in Melo, 1997), stresa (Roberti, Storch in Bravata, 2004), preživljanja prostega časa (Gilchrist, Povey, Dickinson in Povey, 1995), motoričnih sposobnosti (Zarevski, Marušić, Zolotić, Bunjevac in Vukosav, 1998). Za merjenje poteze je bilo razvitih več različnih inštrumentov (Aluja, Garcia in Garcia, 2003a; Arnett, 1994; Hoyle, Stephenson, Palmgreen, Lorch in Donohew, 2002; Zuckerman, 1994), najpogosteje pa se uporablja Zuckermanova lestvica SSS-V, ki z majhnimi popravki izvirne verzije (Zuckerman, Eysenck in Eysenck, 1978) še vedno kaže ustrezne merske lastnosti (Zuckerman, 2007).

Motivacija za napovedovanje izraženosti poteze iskanja dražljajev na osnovi vsebinsko širših osebnostnih potez in temperamentnih lastnosti ni nova. Viken, Kline in Rose (2005) so na osnovi Multifazičnega osebnostnega vprašalnika (MMPI) zgradili novo lestvico, ki je vključevala le tiste postavke izvornega instrumenta, ki so najbolje odražale medosebne razlike udeležencev na Zuckermanovi lestvici SSS-V kot kriterijski meri poteze iskanja dražljajev. Nova lestvica je imela dobro konvergentno in diskriminativno veljavnost ter visoko zanesljivost. Prednost take indirektne ocene je predvsem ta, da omogoča oceno izraženosti poteze iskanja dražljajev na osnovi že zbranih podatkov in ne zahteva ponovnega testiranja celega vzorca udeležencev. Tak pristop se zdi uporaben predvsem v začetnih fazah raziskovanja, pri preverjanju osnovnih predpostavk povezanosti med merjenimi konstrukti ter pri tvorjenju novih raziskovalnih domnev.

Cilj naše študije je bil preveriti napovedno moč odločitvenih dreves pri napovedovanju izraženosti poteze iskanja dražljajev na osnovi temeljnih dimenzij osebnosti ter lastnosti temperamenta, zlasti tistih, ki naj bi bile povezane s procesi vzdraženja in budnosti. Specifično nas je tako zanimala napovedna moč dveh različnih modelov dosežka (regresijskega in klasifikacijskega drevesa) na Zuckermanovi Lestvici iskanja dražljajev SSS-V na osnovi izraženosti (i) temeljnih dimenzij osebnosti po Eysenckovem PEN modelu in (ii) lastnosti temperamenta, ki odražajo način delovanja

centralnega živčnega sistema po Pavlovu. Ustreznost napovedi modelov odločitvenih dreves smo primerjali s klasičnim statističnim modelom multiple linearne regresije. Poleg tega smo želeli kritično ovrednotiti ustreznost uporabe odločitvenih dreves v širšem okviru psiholoških raziskav.

Metoda

Udeleženci

V raziskavi je sodeloval 201 udeleženec: 136 žensk (M = 34,91 let, SD = 11,79 let) in 65 moških (M = 35,49 let, SD = 11,15 let). Vzorčenje je potekalo po metodi snežene kepe; začetno seme je sestavljalo 25 študentov 4. letnika psihologije na Univerzi v Ljubljani (v študijskem letu 2004/2005), ki so na osnovi svojih socialnih mrež in poznanstev pomagali oblikovati končni vzorec udeležencev. Po izobrazbeni strukturi in socialno-ekonomskem položaju je bil vzorec heterogen. Predpostavk o reprezentativnosti vzorca nismo preverjali. Vsi udeleženci so bili na dan testiranja starejši od 18 let.

Pripomočki

Lestvica iskanja dražljajev

Lestvica iskanja dražljajev SSS-V (Zuckerman, 1994) meri štiri dimenzije iskanja dražljajev: iskanje vznemirjenja in pustolovščin (TAS; angl. *Thrill and adventure seeking*), iskanje doživetij (ES; angl. *Experience seeking*), dezinhibicijo (Dis; angl. *Disinhibition*) in nagnjenost k dolgočasju (BS; angl. *Boredom susceptibility*). Inštrument vključuje 40 parov alternativnih trditev, med katerimi posameznik izbira po metodi prisilne izbire. Kljub odgovarjanju po načelu prisilne izbire rezultati niso ipsativni, kar omogoča analizo medosebnih razlik v dosežkih (Zuckerman, prav tam). Skupni dosežek ocenimo na osnovi seštevka dosežkov po posameznih dimenzijah.

V slovenskem prostoru je bila do nedavnega prevedena in uporabljana le starejša Zuckermanova lestvica SSS-IV (Lamovec,1988), zato je bila za potrebe raziskave opravljena jezikovna priredba izvirne angleške verzije (Zuckerman, 1994). Neodvisno je prevod opravil tudi ekspert za angleški jezik, nato pa je profesor slovenskega jezika preveril še ustreznost postavk lestvice z vidika njihove vsebinske, skladenjske in pravopisne ustreznosti. Končna verzija vprašalnika je bila oblikovana v soglasju z obema ekspertoma. Prirejena lestvica SSS-V je dostopna pri avtorju.

Eysenckov osebnostni vprašalnik

Eysenckov osebnostni vprašalnik (Eysensk in Eysenck, 2003) meri tri temeljne dimenzije osebnosti: ekstravertnost (E), nevroticizem (N) in psihoticizem (P). Do-

datno je vključena tudi lestvica L, ki meri tendenco k dajanju socialno zaželenih odgovorov. Kratka verzija vprašalnika vsebuje 48 postavk, na katere posameznik odgovarja z lestvico tipa da-ne. Koeficienti zanesljivosti se na slovenskem standardizacijskem vzorcu gibljejo v razponu od $\alpha = 0{,}63$ do $\alpha = 0{,}86$.

Strelauov vprašalnik temperamenta po Pavlovu

Strelauov vprašalnik temperamenta po Pavlovu (Bucik, 2000) nudi informacijo o treh glavnih dimenzijah temperamenta, skladno s Strelauovim modelom Pavlovove teorije lastnosti centralnega živčnega sistema: moči ekscitacije (ME), moči inhibicije (MI) in mobilnosti živčnega sistema (MO). Inštrument sestavlja 60 postavk, na katere posameznik odgovarja na 4-stopenjski lestvici. Standardizacija slovenske verzije v času testiranja še ni bila dokončana (Zadravec, 2000), vendar so koeficienti zanesljivosti, evalvirani na osnovi delovne verzije vprašalnika, potrjevali dobro zanesljivost tudi v domači priredbi inštrumenta ($\alpha_{ME} = 0{,}82$; $\alpha_{MI} = 0{,}75$; $\alpha_{MO} = 0{,}87$).

Postopek

Testiranje je bilo izpeljano januarja 2005 po vnaprejšnjem dogovoru z vsakim udeležencem. Testiranje je potekalo individualno; poleg avtorja sta pri izvedbi testiranja sodelovali še dve univerzitetni diplomirani psihologinji. Vsi udeleženci so bili seznanjeni z namenom raziskave in so sodelovali prostovoljno. Testiranje je v povprečju trajalo 45 minut, pri čemer čas izpolnjevanja vprašalnikov ni bil omejen.

Statistična analiza

Za statistične izračune, gradnjo napovednih modelov in njihovo vizualizacijo smo uporabili programsko okolje R za statistično analizo in grafiko (R Development Core Team, 2007). Dimenzionalnost Lestvice iskanja dražljajev SSS-V smo preverili z analizo glavnih komponent z uporabo funkcije *prcomp*. Za lažjo interpretacijo glavnih komponent smo uporabili Varimax rotacijo.

Napovedne modele smo gradili z uporabo paketa RWeka. Multipli regresijski model smo zgradili s klicem metod razreda LinearRegression, odločitvena drevesa pa s klicem metod v razredih M5P in J48, kjer smo metode prvega razreda uporabili za gradnjo regresijskega, metode drugega razreda pa za gradnjo klasifikacijskega drevesa. Diskretizacijo zvezne odvisne spremenljivke smo za potrebe algoritma J4.8 opravili na osnovi sheme enakih frekvenc po intervalih (Dougherty, Kohavi in Sahami, 1995). Število razredov odvisne spremenljivke smo nato izbrali na osnovi optimizacije klasifikacijske točnosti. Pri vseh uporabljenih algoritmih strojnega učenja smo uporabili privzete vrednosti parametrov. Za ocenjevanje moči povezanosti med

dejanskimi in napovedanimi vrednostmi odvisne spremenljivke smo pri modelih multiple regresije in regresijskega drevesa uporabili Pearsonov koeficient korelacije, pri klasifikacijskem drevesu pa zaradi diskretnega tipa rešitve Cohenov κ koeficient. Poleg tega smo za oceno prileganja podatkov modelom uporabili še dve različni meri točnosti napovedi (Witten in Frank, 2005): srednjo absolutno napako (*MAE*) in koren relativne srednje kvadratne napake (*RRSE*). Prva ocenjuje absolutno odstopanje napovedi od dejanskih vrednosti, njena velikost pa je odvisna od dejanskega razpona možnih vrednosti funkcije. Ta problem rešimo z uporabo relativne mere *RRSE*, ki ocenjuje relativno uspešnost napovedi v primerjavi s povprečno vrednostjo odvisne spremenljivke na testni množici podatkov in je za prostor sprejemljivih hipotez $RRSE < 1$.

Veljavnost posameznih napovednih modelov smo preizkusili tako, da smo posamezen model najprej zgradili na osnovi učne množice podatkov in ga nato preizkusili na množici testnih podatkov. Povprečno veljavnost smo ocenili na osnovi 10-kratnega prečnega preverjanja, kjer množico vseh primerov slučajno razdelimo na 10 disjunktnih podmnožic (približno) enake velikosti; 9/10 primerov uporabimo kot učno množico, na osnovi katere zgradimo napovedni model, preostala 1/10 primerov pa nam služi kot testna množica za izračun ocene klasifikacijske točnosti. Postopek ponovimo desetkrat, skupno oceno klasifikacijske točnosti pa izračunamo kot povprečje vseh desetih ocen (Witten in Frank, 2005).

Rezultati

Opisne statistike in analiza dimenzionalnosti lestvice SSS-V

Rudarjenje po podatkih smo začeli s pregledom osnovnih statistik, mer povezanosti med spremenljivkami in preverjanjem dimenzionalnosti lestvice SSS-V. Opisne statistike vzorca so zbrane v tabeli 1.

Razlike med spoloma na proučevanih spremenljivkah so izražene s Cohenovim *d*, kjer vrednost cenilke $d = 0{,}2$ predstavlja majhen, $d = 0{,}5$ srednji in $d = 0{,}8$ velik učinek. Ob upoštevanju kriterija praktične pomembnosti se pomembne razlike med spoloma kažejo na skupnem dosežku lestvice SSS-V ter podlestvicah TAS in Dis. Z izjemo lestvice P ter podlestvic ES in BS je zanesljivost merjenja zadovoljiva. O nizkih koeficientih zanesljivosti lestvice psihoticizma poroča večina študij, ki so proučevale metrične lastnosti inštrumenta EPQ (glej npr. Caruso, Witkiewitz, Belcourt-Dittloff in Gottlieb, 2001; Lewis, Francis, Shevlin in Forrest, 2002). Koeficienti zanesljivosti podlestvic SSS-V so primerljivi s koeficienti, o katerih poroča Zadravec (2003; TAS = 0,78; ES = 0,54; Dis = 0,69; BS = 0,53; SSS = 0,81). Šibka zanesljivost podlestvic lestvice ES in BS ne preseneča, saj tudi drugi avtorji poročajo o podobnih lastnostih

Tabela 1. *Osnovne statistike vzorca.*

	M	*SD*	$M_{moški}$	$SD_{moški}$	$M_{ženske}$	$SD_{ženske}$	*d*	α
E	8,32	3,15	8,22	3,07	8,38	3,19	–0,05	0,84
N	5,68	3,46	5,20	3,64	5,91	3,37	–0,20	0,85
P	3,52	1,87	3,69	2,07	3,43	1,78	0,13	0,46
L	5,03	2,98	4,35	3,06	5,35	2,89	–0,34	0,78
ME	49,52	7,32	51,54	7,22	48,56	7,20	0,41	0,80
MI	55,36	7,73	54,92	8,64	55,57	7,28	–0,08	0,81
MO	58,02	7,64	58,06	7,13	58,00	7,90	0,01	0,85
TAS	4,64	3,02	5,97	2,95	4,01	2,85	0,68	0,82
ES	5,66	1,99	5,46	2,02	5,75	1,98	–0,14	0,52
Dis	4,75	2,49	5,71	2,56	4,29	2,33	0,58	0,72
BS	2,67	1,82	3,05	1,93	2,49	1,75	0,30	0,48
SSS	17,71	6,80	20,18	6,56	16,53	6,62	0,55	0,83

Opombe: d = Cohenova *d* statistika, α = Cronbachov koeficient zanesljivosti. Za oznake spremenljivk glej besedilo.

inštrumenta (Aluja, Garcia in Garcia, 2003b, Zuckerman, 1994). Mere povezanosti med spremenljivkami so povzete v tabeli 2.

Dimenzionalnost lestvice SSS-V smo preverili z analizo glavnih komponent. Matrika nasičenosti z vsiljeno štirikomponentno strukturo je prikazana v tabeli 3. Štiri komponente pojasnijo dobrih 30 % celotne variabilnosti, kar je razmeroma malo. S prvo komponento so močno nasičene postavke podlestvice TAS, z drugo komponento pa postavke podlestvice Dis. Medtem ko podlestvici TAS in Dis kažeta razmeroma enoznačen vzorec nasičenosti, vsaka s svojo komponento, je struktura ostalih dveh komponent manj jasna. Očitno je, da so nizke ocene zanesljivosti posledica tega, da podlestvici ES in BS nista dovolj homogeni, kar se kaže v razpršenem vzorcu nasičenosti. Postavke podlestvice ES so povezane z vsemi štirimi komponentami, postavke podlestvice BS pa predvsem z drugo in tretjo komponento. Rezultati sicer govorijo v prid večkomponentnemu modelu iskanja dražljajev, vendar smo se zaradi nejasne strukture nasičenosti odločili, da kot mero izraženosti poteze iskanja dražljajev definiramo skupni dosežek na lestvici SSS-V. Tako več- kot enokomponentni model iskanja dražljajev bi sicer kazalo preveriti tudi s konfimatornim pristopom, vendar bi to močno preseglo obseg in namen prispevka.

Napovedni modeli dosežka na Lestvici SSS-V

Napovedne modele smo gradili na osnovi podatkovnega okvirja, ki ga je sestavljalo osem spremenljivk: ekstravertnost (E), nevroticizem (N), psihoticizem (P), moč ekscitacije (ME), moč inhibicije (MI), mobilnost živčnega sistema (MO), spol (SPOL) in odvisna spremenljivka SSS. Zaradi parsimoničnosti rešitve ter problema

Tabela 2. *Korelacijska matrika*

	E	N	P	L	ME	MI	MO	TAS	ES	Dis	BS	SSS
E		–0,22	0,02	–0,06	0,30	0,04	**0,47**	0,25	0,14	0,37	0,18	0,35
N	–0,22		0,34	–0,07	**–0,64**	–0,37	–0,34	–0,21	–0,16	–0,04	0,02	–0,15
P	**0,30**	**0,31**		–0,33	–0,11	**–0,59**	0,00	0,12	0,22	0,32	**0,43**	0,37
L	–0,20	–0,18	**–0,43**		–0,01	0,26	–0,19	–0,27	–0,27	**–0,41**	–0,23	**–0,43**
ME	0,26	**–0,45**	0,06	–0,07		0,26	**0,60**	**0,41**	0,26	0,14	0,01	0,33
MI	–0,10	**–0,37**	**–0,43**	**0,43**	**0,31**		0,04	–0,03	0,00	–0,28	**–0,49**	–0,26
MO	**0,51**	–0,25	0,15	–0,11	**0,58**	0,12		0,27	0,18	0,11	0,04	0,23
TAS	0,13	0,04	**0,35**	**–0,35**	**0,34**	–0,15	**0,34**		0,35	0,35	–0,05	**0,68**
ES	0,17	–0,09	**0,30**	**–0,31**	**0,44**	0,06	**0,32**	**0,59**		**0,48**	0,16	**0,70**
Dis	**0,36**	0,11	**0,39**	**–0,39**	0,18	–0,26	**0,31**	**0,43**	**0,37**		**0,48**	**0,84**
BS	0,20	0,12	**0,38**	**–0,30**	0,10	–0,27	0,15	0,25	0,28	**0,37**		**0,51**
SSS	**0,28**	0,06	**0,48**	**–0,46**	**0,37**	–0,21	**0,40**	**0,83**	**0,76**	**0,75**	**0,59**	

Opombe: Zgornji korelacijski trikotnik se nanaša na moške, spodnji na ženske. Poudarjene so vrednosti korelacijskih koeficientov, za katere velja $p < ,01$. Za oznake spremenljivk glej besedilo.

Tabela 3. *Nasičenosti za štiri glavne komponente, rotirane z Varimax rotacijo*

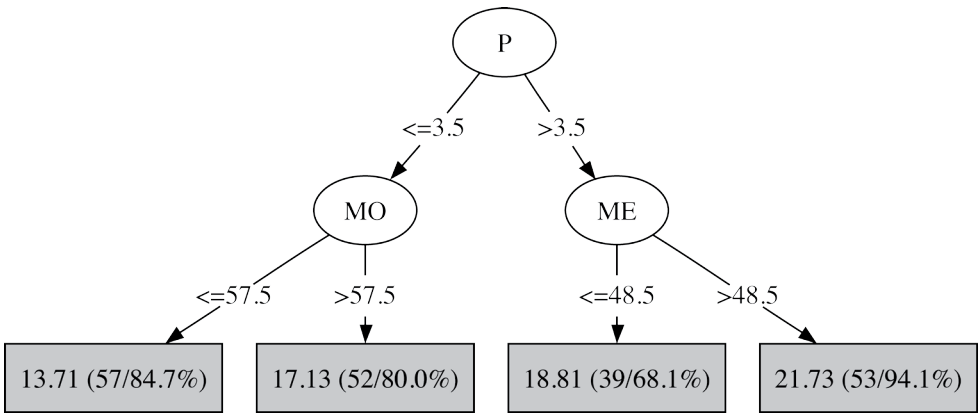
Postavka	*Komp 1*	*Komp 2*	*Komp 3*	*Komp 4*
TAS1	0,51	0,02	0,17	0,04
TAS2	0,50	0,19	0,28	0,29
TAS3	0,73	–0,01	–0,18	–0,05
TAS4	0,77	0,02	–0,12	0,00
TAS5	0,63	0,05	0,07	–0,08
TAS6	0,58	0,01	0,08	0,07
TAS7	0,70	0,06	0,07	0,15
TAS8	0,53	0,22	0,09	0,09
TAS9	0,45	0,17	0,06	0,16
TAS10	0,49	0,04	0,18	0,27
ES1	0,07	–0,10	–0,09	0,30
ES2	0,45	0,07	0,39	0,01
ES3	0,27	0,11	0,18	0,34
ES4	0,21	0,21	0,47	0,09
ES5	0,10	–0,08	0,13	0,15
ES6	0,24	–0,09	0,55	0,18
ES7	0,09	0,08	0,23	0,44
ES8	–0,06	–0,13	0,16	0,61
ES9	0,01	0,07	–0,10	0,51
ES10	0,11	0,10	0,00	0,35
Dis1	0,13	0,59	0,14	0,00
Dis2	0,37	0,32	–0,12	0,19
Dis3	0,05	0,35	0,09	0,48
Dis4	0,42	0,15	0,44	0,06
Dis5	0,12	0,54	0,09	0,02
Dis6	0,15	0,60	0,16	–0,08
Dis7	0,03	0,37	0,19	0,27
Dis8	0,19	0,21	0,09	0,28
Dis9	0,26	0,50	–0,15	0,09
Dis10	0,01	0,64	0,12	0,00
BS1	–0,05	0,22	0,09	–0,35
BS2	0,08	0,28	0,26	–0,21
BS3	–0,06	0,29	0,21	0,23
BS4	–0,01	0,01	0,48	–0,10
BS5	0,00	0,35	–0,28	–0,09
BS6	0,09	0,25	0,50	0,05
BS7	–0,08	0,06	0,54	0,01
BS8	0,19	0,30	0,05	0,28
BS9	–0,06	0,52	–0,09	0,13
BS10	0,07	0,41	0,13	–0,27
λ	4,52	3,25	2,38	2,29
σ^2	11,31	8,12	5,96	5,74

Opombe: λ – lastna vrednost komponente; σ^2 – odstotek pojasnjene variance.

kolinearnosti prediktorjev smo iz napovedi izločili vse podlestvice iskanja dražljajev in lestvico lažnivosti. Ker je bila na skupnem dosežku lestvice SSS-V razlika med spoloma praktično pomembna, smo v napovedne modele vključili tudi diskretno spremenljivko SPOL. Spremenljivke starost v analizo nismo vključili.

Regresijsko drevo

Napovedni model, zgrajen na osnovi induktivnega učenja regresijskega drevesa, je prikazan na sliki 1.



Slika 1. Regresijsko drevo. Vsak od listov predstavlja sklepni del pravila, ki preslika vrednosti neodvisnih spremenljivk v vrednost regresijske spremenljivke. V oklepaju je navedeno število učnih primerov, ki zadoščajo danemu pravilu in indeks prileganja podatkov modelu (RRSE / MAE). Za oznake spremenljivk glej besedilo.

Kot najbolj informativna spremenljivka za napovedovanje skupnega dosežka SSS se je izkazala dimenzija psihoticizma (P), ki ji ob prvi členitvi drevesa sledita mobilnost živčnega sistema (MO) ter moč ekscitacije (ME). Posamezniku z visokim dosežkom na dimenzijah psihoticizma ($> 3{,}5 \approx 4$) ter moči ekscitacije ($> 48{,}5 \approx 49$) bomo pripisali 21,73 (≈ 22) točke na skupnem dosežku na lestvici SSS, posamezniku z nizkim psihoticizmom ($\leq 3{,}5 \approx 4$) in visoko mobilnostjo živčnega sistema ($> 57{,}5 \approx 58$) pa 17,13 (≈ 18) točk. Točnost napovedi odvisne spremenljivke, ocenjena na osnovi korelacijskega koeficienta med empiričnimi in napovedanimi vrednostmi na polni učni množici podatkov, znaša $r = 0{,}55$ ($p < 0{,}001$; $r^2_{popr} = 0{,}30$), po 10-kratnem prečnem preverjanju veljavnosti pa se korelacijski koeficient zniža na $r = 0{,}40$ ($p < 0{,}001$; $r^2_{popr} = 0{,}16$). S tremi prediktorji tako pojasnimo 16 % variance skupnega dosežka SSS.

Klasifikacijsko drevo

Klasifikacijsko drevo za razliko od regresijskega drevesa napoveduje diskretno spremenljivko. V našem primeru smo zato skupni dosežek na lestvici SSS najprej diskretizirali. Klasifikacijska točnost modela po 10-kratnem prečnem preverjanju v odvisnosti od števila razredov odvisne spremenljivke je prikazana v tabeli 4. Kot optimalna se je izkazala rešitev z razbitjem zveznega dosežka na dva razreda, s katerima dosežemo relativno maksimalno klasifikacijsko točnost. Klasifikacijsko drevo v tej obliki je prikazano na sliki 2.

Tabela 4. *Klasifikacijska točnost algoritma J4.8 glede na število razredov skupnega dosežka SSS-V*

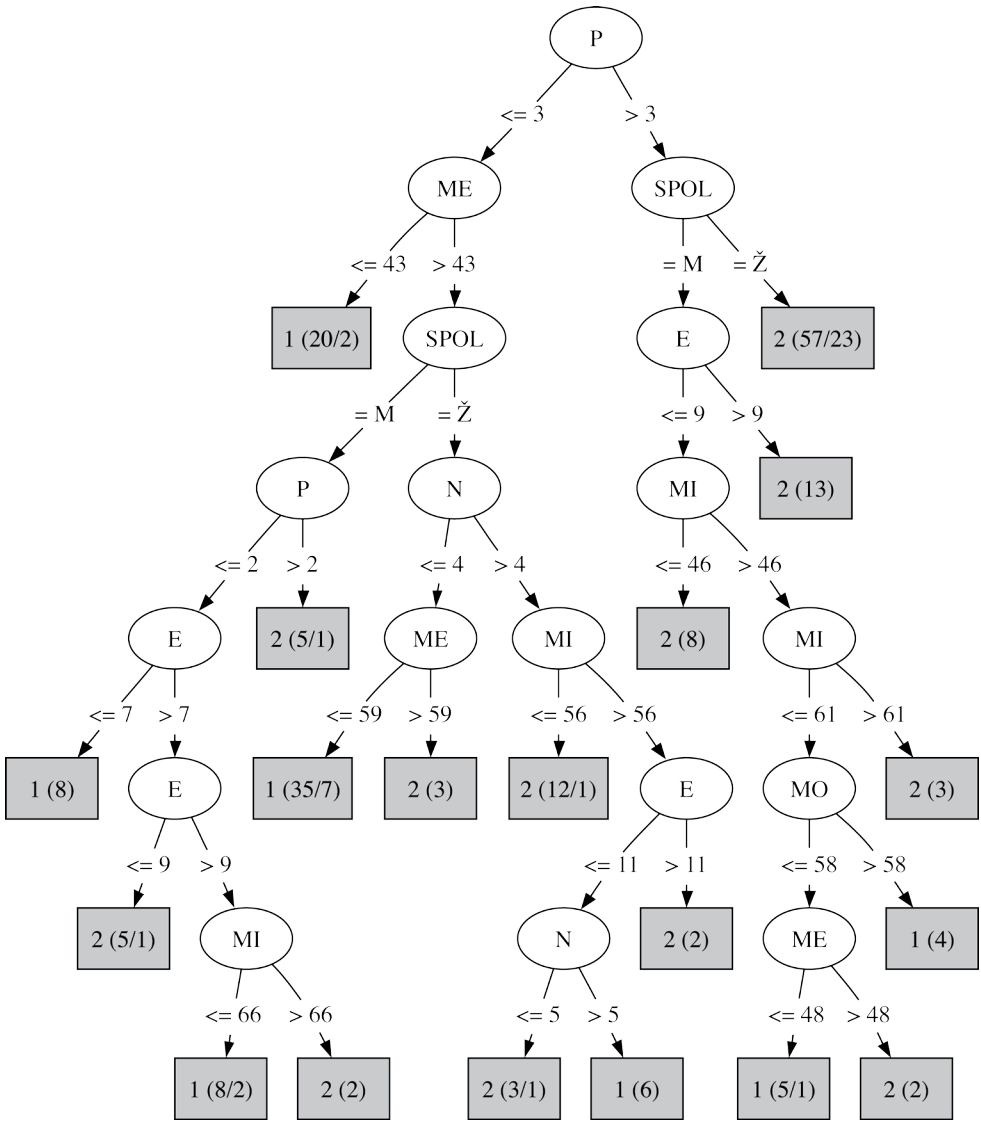
n	KT	*MAE*	*RRSE*
2	0,58	0,45	1,14
3	0,44	0,38	1,20
4	0,31	0,35	1,24
5	0,35	0,28	1,16
6	0,30	0,24	1,18
7	0,17	0,23	1,23
8	0,14	0,21	1,24
9	0,21	0,18	1,19
10	0,11	0,18	1,26

Opombe: n = št. razredov, KT = klasifikacijska točnost, *MAE* = srednja absolutna napaka, *RRSE* = koren relativne srednje kvadratne napake.

Tudi v tem primeru je najbolj informativna spremenljivka psihoticizem (P), ki ji sledita moč ekscitacije (ME) in spol (SPOL). Posameznica (= Ž) z visokim dosežkom na psihoticizmu (> 3) bo tako uvrščena v razred 2 (visok dosežek na lestvici SSS), posameznik (bodisi moški bodisi ženska) z nizkim psihoticizmom (≤ 3) in nizko stopnjo moči ekscitacije živčnega sistema pa v razred 1 (nizek dosežek na lestvici SSS). Točnost napovedi tako induciranega modela, ocenjena na osnovi κ koeficienta na polni učni množici podatkov znaša $\kappa = 0{,}61$ ($p = 0{,}00$), po 10-kratnem prečnem preverjanju pa se koeficient zniža na $\kappa = 0{,}16$ ($p = 0{,}02$).

Linearni regresijski model

Za gradnjo linearnega regresijskega modela smo uporabili metodo po korakih. Število prediktorjev smo določili z minimizacijo AIC funkcije. AIC funkcija (angl. *Akaike Information Criterion*) je mera prileganja podatkov statističnemu modelu, osnovana na konceptu entropije (Hastie, Tibshirani in Friedman, 2001; Venables in



Slika 2. Klasifikacijsko drevo. Vsak od listov predstavlja sklepni del pravila, ki preslika vrednosti neodvisnih spremenljivk v ustrezen razred (1 = nizek dosežek, 2 = visok dosežek). V oklepaju je navedeno število pravilno / nepravilno razvrščenih primerov. Za oznake spremenljivk glej besedilo.

Ripley, 2002). Rezultati končnega modela so povzeti v tabeli 5.

Višja kot je absolutna vrednost standardiziranega regresijskega koeficienta β, višja je relativna pomembnost prediktorja glede na ostale prediktorje. Na skupni

Tabela 5. *Multipli regresijski model*

	B	*SE(B)*	β	*t*	*p*
E	0,27	0,13	0,12	2,09	0,04
P	1,17	0,24	0,32	4,88	0,00
ME	0,34	0,06	0,36	5,78	0,00
MI	–0,14	0,06	–0,16	–2,31	0,02
SPOL	–2,30	0,83	–0,16	–2,77	0,01

Opombe: za oznake spremenljivk glej besedilo.

dosežek na lestvici SSS statistično značilno vplivajo ekstravertnost (E), psihoticizem (P), moč ekscitacije (ME), moč inhibicije (MI) in spol (SPOL). Dosežek na lestvici povečujejo ekstravertnost, psihoticizem in moč ekscitacije, zmanjšuje pa moč inhibicije. Moški bodo imeli v povprečju višjo napoved kot ženske. Regresijsko-diagnostični postopki so potrdili, da je model ustrezen. Model se je pokazal kot statistično značilno boljši od ničelnega; $F(5, 195) = 25{,}28$; $p < 0{,}001$. Tudi s kolinearnostjo ni bilo težav, saj so bile vse vrednosti VIF s sprejemljivem razponu ($1{,}06 < \text{VIF} < 1{,}49$). Predpostavka o normalnosti porazdelitve ostankov je bila za predlagani regresijski model izpolnjena. S petimi prediktorji pojasnimo slabih 40 % variance dosežka SSS ($R = 0{,}63$; $R^2_{\text{popr}} = 0{,}38$). Po 10-kratnem prečnem preverjanju se napovedna moč modela nekoliko zniža ($R = 0{,}58$; $R^2_{\text{popr}} = 0{,}32$). Analizo smo ponovili še s tremi prediktorji, in sicer tistimi, na osnovi katerih je bilo zgrajeno regresijsko drevo (P, ME in MO). S temi tremi prediktorji smo na polni učni množici pojasnili 34 % celotne variance dosežka na lestvici ($R = 0{,}59$; $R^2_{\text{popr}} = 0{,}34$), po 10-kratnem prečnem preverjanju pa 30 % celotne variance ($R = 0{,}56$; $R^2_{\text{popr}} = 0{,}30$).

Primerjava modelov

Vsi uporabljeni algoritmi so privzeto določili zadostno število prediktorskih spremenljivk za napovedovanje oz. razvrščanje učnih primerov. Najkompleksnejše je klasifikacijsko drevo s šestimi prediktorji, ki skupaj tvorijo 19 odločitvenih pravil. Model multiple regresije je sestavljen iz šestih prediktorjev, regresijsko drevo pa iz treh prediktorjev, ki skupaj tvorijo štiri odločitvena pravila. Tako glede na mere povezanosti med dejanskimi in napovedanimi vrednostmi kot glede na različne indekse prileganja podatkov modelom (tabela 6), se je kot najuspešnejši izkazal model multiple regresije. Model multiple regresije se je kot najboljši izkazal tudi v napovednem modelu z vsiljenimi tremi prediktorji regresijskega drevesa ($MAE = 4{,}58$; $RRSE = 0{,}83$). Indeks $RRSE > 1$ v primeru klasifikacijskega drevesa nakazuje, da lahko boljšo rešitev dosežemo tudi s trivialno funkcijo (npr. aritmetično sredino).

Tabela 6. *Primerjava napovednih modelov.*

	MLR	M5P	J48
MAE	4,63	5,16	0,45
RRSE	0,81	0,91	1,14

Opombe: MLR = multipla linearna regresija, M5P = regresijsko drevo, J48 = klasifikacijsko drevo, MAE = srednja absolutna napaka, RRSE = koren relativne srednje kvadratne napake.

Razprava

Za izdelavo napovednega modela skupnega dosežka na Lestvici iskanja dražljajev SSS-V smo uporabili dva različna tipa odločitvenih dreves, in sicer regresijsko ter klasifikacijsko drevo. Dobljene rezultate smo primerjali z rezultati klasične multiple linearne regresije. Poleg tega, kako točno posamezen model napoveduje vrednosti učnih primerov, nas je pri oceni zanesljivosti modelov zanimala predvsem njihova točnost oz. napovedna moč pri napovedovanju vrednosti neznanih primerov. Točnost napovedi modela na neznanih primerih smo ocenili s prečnim preverjanjem in ugotovili, da je kvaliteta zgrajenih modelov močno odvisna od uporabljene metode. Kot napovedno najmočnejši se je izkazal model multiple linearne regresije. Klasifikacijska točnost regresijskega drevesa je bila sicer nekoliko nižja, vendar je potrebno upoštevati, da model vključuje le tri prediktorje, za razliko od napovedno močnejšega modela multiple regresije s petimi prediktorji. Klasifikacijsko drevo se je za napovedovanje izkazalo kot neuporabno.

Empirični izsledki iz različnih problemskih domen (Chae, Ho, Cho, Lee in Ji, 2001; Delen, Walker in Kadam, 2005) potrjujejo, da dobimo pri velikih učnih množicah z različnimi algoritmi za gradnjo odločitvenih dreves običajno natančnejše napovedne modele kot s klasičnimi linearnimi regresijskimi modeli. Slednji se nasprotno bolje obnesejo pri manjših učnih množicah (Young Soo, 2008; Ženko in Džeroski, 2002). Kljub temu pa odločitvena drevesa, zgrajena na osnovi manjšega števila primerov, niso neuporabna. Glavno prednost odločitvenih dreves na prikazanem primeru gre iskati predvsem v smeri hitrega, razumljivega in grafično enostavnega prikaza odnosa med posameznimi merjenimi spremenljivkami. Odločitveno drevo predstavlja klasifikacijsko funkcijo (podobno kot regresijska enačba), ki pa je hkrati simboličen opis podatkov in povzetek zakonitosti v danem naboru podatkov. Obstoječi načini opisa podatkov s korelacijskimi matrikami in regresijskimi enačbami so zlasti pri velikem številu spremenljivk nepregledni in često se zgodi, da raziskovalec v množici korelacijskih koeficientov in konstant ne more ločiti »gozda od dreves«. Prednost odločitvenih dreves pred regresijsko analizo se pokaže tudi v primerih, ko so prediktorji v interakciji. Slednjo je namreč z običajnimi regresijskimi metodami precej težko odkrivati oz. dobro modelirati. Odločitvena drevesa omogočajo tudi dobro vizualizacijo podatkov. Vizualizacija podatkov se je namreč

z razmahom podatkovnega rudarjenja otresla priokusa nepotrebnega okrasja in postala ne samo nujna spremljevalka sodobnega znanstvenega poročanja, pač pa tudi pomembna znanstvena disciplina (Wilkinson, 2005). Vizualizacija podatkovnih struktur in rezultatov statističnih analiz je tako sestavni del ali pa celo končni cilj številnih statističnih orodij in metod. Slabša napovedna moč zato po našem mnenju ne odtehta preprostosti in razumljive grafične slovnice modelov odločitvenih dreves, zlasti ne v fazah eksploratornega rudarjenja po podatkih, preiskovanja strukture podatkov in kreativnega snovanja novih raziskovalnih domnev. Poleg tega je indukcija odločitvenih dreves bistveno hitrejša in z vidika povprečnega uporabnika manj zahtevna kot uporaba linearne regresije, ki kljub močni programski podpori, zlasti v fazi interpretacije modela, zahteva več znanja in izkušenj.

Predstavljeni podatki so v prvi vrsti služili ilustraciji podatkovnega rudarjenja, kljub temu pa lahko na njihovi osnovi potrdimo nekatere že znane odnose med potezo iskanja dražljajev, temeljnimi dimenzijami osebnosti in lastnostmi temperamenta. Zgrajeni napovedni modeli nakazujejo, da lahko variabilnost poteze iskanja dražljajev pojasnimo predvsem z medosebnimi razlikami v psihoticizmu in nekaterih lastnostih temperamenta. Psihoticizem se je v vseh treh modelih izkazal kot najbolj informativna spremenljivka, ki ji sledita bodisi moč inhibicije bodisi mobilnosti živčnega sistema (regresijsko drevo) oz. moč ekscitacije (regresijsko in klasifikacijsko drevo). Podatki govorijo v prid Zuckermanovi (1994) predpostavki o povezanosti iskanja dražljajev z dimenzijo psihoticizma in postavljajo pod vprašaj Eysenckovo trditev (Glicksohn in Abulafia, 2001; Zuckerman idr., 1978) o dominantni povezavi iskanja dražljajev z ekstravertnostjo. Ekstravertnost sicer nastopa tako v modelu multiple regresije kot v klasifikacijskem drevesu, vendar se kot prediktor po napovedni moči ne more primerjati s psihoticizmom. Visoka moč ekscitacije centralnega živčevja pri posameznikih z močno izraženo potezo iskanja dražljajev se nanaša tako na sposobnost učinkovitega delovanja pod vplivom močnih in neprijetnih dražljajev kot na različne zavestno izbrane aktivnosti, ki neposredno vplivajo na povečano aktivacijo centralnega živčevja (Strelau in Zawadski, 1997). Po drugi strani pa zaradi šibkih inhibitornih procesov težje vzpostavljajo stanje pogojne inhibicije (Zuckerman, 2005). Mobilnost živčnega sistema kot prediktor iskanja dražljajev nastopa le pri regresijskem drevesu. Navezuje se na sposobnost hitrega reagiranja na spremembe v okolju in v tem kontekstu lahko sprejmemo Eysenckovo predpostavko, da je višja optimalna raven budnosti pri posameznikih z višje izraženo potezo iskanja dražljajev povezana z večjo mobilnostjo centralnega živčnega sistema (Zuckerman, 2005). Nevronski sistem posameznikov z nizko stopnjo habitualnega vzburjenja se hitro adaptira na znane dražljaje ter zato aktivno izbira tiste aktivnosti, ki mu zagotavljajo konstanten dotok novih dražljajev. Tako je omogočena stalna avtoregulacija optimalnega nivoja vzburjenja.

Nikakor ni nujno, da je predstavljen način analize podatkov za dano problemsko domeno najboljši. Odločitvena drevesa smo izbrali zato, ker predstavljajo optimalno razmerje med računsko zahtevnostjo, razumljivostjo in elegantnostjo

grafičnega prikaza. Med pisano paleto sodobnih metod rudarjenja po podatkih bi zato zlahka našli take, pri katerih bi bila natančnost napovedovanja veliko večja (npr. metoda podpornih vektorjev; Wu idr., 2008). Pokazalo se je tudi, da je transformacija razmernostne spremenljivke v diskretno nesmiselna, kljub temu da smo implicitno predpostavljali, da bomo s diskretizacijo odvisno spremenljivko očistili neželenega šuma. Nadaljnje raziskovanje uporabe metod podatkovnega rudarjenja na področju psihologije zahteva veliko bolj dodelan eksperimentalni načrt, vključitev podatkov na različnih merskih lestvicah, preizkus različnih algoritmov, njihovo statistično validacijo itd. Posebno pozornost gre posvetiti odnosu med informativnostjo oz. napovedno močjo spremenljivk in njihovo zanesljivostjo. V našem eksperimentu se je tako npr. pokazalo, da je zelo slabo zanesljiva lestvica psihoticizma nastopala kot pomemben prediktor v vseh treh evalviranih modelih. Na vsa ta vprašanja bo potrebno odgovoriti v prihodnjih raziskavah.

Odločitvena drevesa imajo s psihologijo veliko več skupnega, kot se zdi na prvi pogled. Kategorizacija oz. klasifikacija je temelj zaznave, mišljenja, jezika in aktivnosti (Augoustinos in Walker, 1996). Bruner, Goodnow in Austin (1956) pravijo, da je klasifikacija ena od najbolj osnovnih in splošnih oblik spoznavanja. Njena glavna funkcija je redukcija kompleksnih dražljajev, kar omogoča vzpostavljanje reda med dražljaji ter učinkovitejšo komunikacijo v okolju. S pojavom prvih računskih strojev so se začeli tudi zametki strojno podprte klasifikacije. Klasifikacija danes predstavlja paradno metodo podatkovnega rudarjenja, strojnega učenja in umetne inteligentnosti. Temeljni kamen, na osnovi katerega je računalniška znanost razvijala svoje klasifikacijske algoritme, predstavlja že omenjeno delo treh kognitivnih psihologov Brunerja, Goodnowa in Austina (1956) z naslovom »A study of thinking«, v katerem so človekov proces klasifikacije opisali s formalnim teoretičnim jezikom ter s tem omogočili izgradnjo prvih strojnih algoritmov. Področje odkrivanja zakonitosti iz podatkov, vključno z metodami podatkovnega rudarjenja, strojnega učenja in umetne inteligentnosti si zato že zaradi tradicije zasluži večjo, tako teoretično kot aplikativno, pozornost raziskovalcev na področju psihologije.

Zahvala

Avtor se iskreno zahvaljuje obema anonimnima recenzentoma, ki sta s svojimi konstruktivnimi pripombami izboljšala kakovost prispevka.

Literatura

Aluja, A., Garcia, O. in Garcia, L. F. (2003a). Psychometrics properties of the Zuckerman-Kuhlman Personality Questionnaire (ZKPQ-III-R): A study of a shortened form. *Personality and Individual Differences, 34*(7), 1083–1097.

Aluja, A., Garcia, O. in Garcia, L. F. (2003b). Relationships among extraversion, open-

ness to experience, and sensation seeking. *Personality and Individual Differences, 35*(3), 671–680.

Andrew, M. in Cronin, C. (1997). Two measures of sensation seeking as predictors of alcohol use among high school males. *Personality and Individual Differences, 22*(3), 393–401.

Arnett, J. (1994). Sensation seeking: A new conceptualization and a new scale. *Personality and Individual Differences, 16*(2), 289–296.

Augoustinos, M. in Walker, I. (1996). *Social cognition*. London: SAGE.

Baker, R. S. J. d., Barnes, T. in Beck, J. E. (ur.). (2008). *Educational Data Mining 2008: Proceedings of The 1ſst International conference on educational data mining.* Montreal: University of Quebec.

Bardo, M. T., Donohew, R. L. in Harrington, N. G. (1996). Psychobiology of novelty and drug seeking behavior. *Behavioural Brain Research, 77*(1–2), 23–43.

Berthold, M. in Hand, D. J. (ur.). (2007). *Intelligent data analysis*. Berlin: Springer.

Bohanec, M. (2006). *Odločanje in modeli* [Decision making and modeling]. Ljubljana: Društvo matematikov, fizikov in astronomov.

Bruner, J., Goodnow, J., in Austin, G. (1956). *A Study of Thinking*. New York, NY: Wiley.

Bucik, V. (2000). Načela priredbe psiholoških testov iz drugih jezikovnih in kulturnih okolij: primer vprašalnika VTP [Principles of the cross-cultural adaptation of an inventory: The case of the Pavlovian Temperament Survey]. *Psihološka obzorja, 9*(3), 67–78.

Caruso, J. C., Witkiewitz, K., Belcourt-Dittloff, A. in Gottlieb, J. D. (2001). Reliability of scores from Eysenck Personality Questionnaire: A reliability generalization study. *Educational and Psychological Measurement, 61*(4), 675–689.

Chae, Y. M., Ho, S. H., Cho, K. W., Lee, D. H. in Ji, S. H. (2001). Data mining approach to policy analysis in a health insurance domain. *International Journal of Medical Informatics, 62*(2–3), 103–111.

Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003.

Delen, D., Walker, G. in Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine, 34*(2), 113–127.

Dougherty, J., Kohavi, R. in Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. V A. Prieditis in S. Russell (ur.), *Machine Learning: Proceedings of the Twelfth International Conference, July 9–12,* (str. 194–202). Tahoe City, CA: Morgan Kaufmann.

Eysenck, H. J. in Eysenck, S. B. G. (2003). *Eysenckove osebnostne lestvice EPQ-R, IVE: priročnik* [Eysenck personality scales EPQ-R, IVE: Manual]. Ljubljana: Center za psihodiagnostična sredstva.

Gasar, S., Bohanec, M. in Rajkovič, V. (2002). Primerjava treh tipov modelov za napovedovanje uspešnosti zaključka šolanja [Comparison of three types of models for the prediction of final academic achievement]. *Psihološka obzorja, 11*(4), 7–24.

Gilchrist, H., Povey, R., Dickinson, A. in Povey, R. (1995). The Sensation Seeking Scale: Its use in a study of the characteristics of people choosing 'adventure holidays'. *Personality and Individual Differences, 19*(4), 513–516.

Glicksohn, J. in Abulafia, J. (2001). Embedding sensation seeking within the big three. *Personality and Individual Differences, 25*(6), 1085–1099.

Hastie, T., Tibshirani, R. in Friedman, J. (2001). *The elements of statistical learning*. New York, NY: Springer.

Hoyle, R. H., Stephenson, M. T., Palmgreen, P., Lorch, E. P. in Donohew, R. L. (2002). Reliability and validity of a brief measure of sensation seeking. *Personality and Individual Differences, 32*(3), 401–414.

Kobal Grum, D., Arnerič, N., Kobal, A. B., Horvat, M., Ženko, B., Džeroski, S. in Osredkar, J. (2004). Emotions and personality traits in former mercury miners. *Psihološka obzorja, 13*(4), 9–31.

Kononenko, I. in Kuhar, M. (2007). *Machine learning and data mining: Introduction to principles and algorithms*. West Sussex: Horwood.

Lamovec, T. (1988). *Priročnik za psihologijo motivacije in emocij* [Psychology of motivation and emotion: Manual]. Ljubljana: Filozofska fakulteta.

Lewis, C. A., Francis, L. J., Shevlin, M. in Forrest, S. (2002). Confirmatory factor analysis of the French translation of the abbreviated form of the Revised Eysenck Personality Questionnaire (EPQR-A). *European Journal of Psychological Assessment*, *18*(2): 180–186.

Mitchell, T. M. (1997). *Machine learning*. New York, NY: McGraw-Hill.

Munafò, M. R., Yalcin, B., Willis-Owen, S. A. in Flint, J. (2008). Association of the dopamine D4 receptor (DRD4) gene and approach-related personality traits: Meta-analysis and new data. *Biological Psychiatry*, *63*(2), 197–206.

Pliner, P. in Melo, N. (1997). Food neophobia in humans: Effects of manipulated arousal and individual differences in sensation seeking. *Physiology & Behavior, 61*(2), 331–335.

R Development Core Team. (2007). *R: A language and environment for statistical computing*. Vienna, Austria. (ISBN 3-900051-07-0)

Roberti, J. W., Storch, E. A. in Bravata, E. A. (2004). Sensation seeking, exposure to psychosocial stressors, and body modifications in a college population. *Personality and Individual Differences, 37*(6), 1167–1177.

Rosnow, R. L. in Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*(10), 1276–1284.

Slivar, B. (2008). Ugotavljanje vzorca stresorjev pri delu učiteljev v povezavi z zadovoljstvom pri delu. *Psihološka obzorja, 17*(3), 93–112.

Stoel, R. D., De Geus, E. J. C. in Boomsma, D. I. (2006). Genetic analysis of sensation seeking with an extended twin design. *Behavior Genetics, 36*(2), 229–237.

Strelau, J. in Zawadski, B. (1997). Temperament and personality: Eysenck's three superfactors as related to temperamental dimensions. V H. Nyborg (ur.), *The scientific study of human nature: Tribute to Hans J. Eysenck at eighty* (str. 68–91). New York, NY: Elsevier.

Venables, W. N. in Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York, NY: Springer.

Viken, R. J., Kline, M. P. in Rose, R. J. (2005). Development and validation of an MMPI-based Sensation Seeking Scale. *Personality and Individual Differences, 38*(3), 619–625.

Wiesbech, G. A., Wodarz, N., Mauerer, C., Thome, J., Jakob, F. in Boening, J. (1996). Sensation seeking, alcoholism and dopamine activity. *European Psychiatry, 11*(2), 87–92.

Wilkinson, L. (2005). *The grammar of graphics* (2nd ed.). New York, NY: Springer.
Witten, I. H. in Frank E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). San Francisco, CA: Morgan Kaufmann.
Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z., Steinbach, M., Hand, D. J. in Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37.
Young Soo, K. (2008). Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size. *Expert Systems with Applications, 34*(2), 1227–1234.
Zadravec, T. (2000). *Impulzivnost z vidika Eysenckove teorije osebnosti in Pavlovove teorije temperamenta* [Impulsivity in terms of Eysenck's personality theory and Pavlov's theory of temperament]. Neobjavljeno diplomsko delo [Unpublished BSc diploma thesis], Univerza v Ljubljani, Filozofska fakulteta, Ljubljana, Slovenija.
Zadravec, T. (2003). *Konstruktna veljavnost dimenzije impulzivnost in njen odnos do drugih dimenzij osebnosti* [Construct validity of impulsivity and it's relationship with other dimensions of personality]. Neobjavljeno magistrsko delo [Unpublished MA thesis], Univerza v Ljubljani, Filozofska fakulteta, Ljubljana, Slovenija.
Zarevski, P., Marušić, I., Zolotić, S., Bunjevac, T. in Vukosav, Ž. (1998). Contribution of Arnett's Inventory of Sensation Seeking and Zuckerman's Sensation Seeking Scale to the differentiation of athletes engaged in high and low risk sports. *Personality and Individual Differences, 25*(4), 763–768.
Zuckerman, M. (1969). Theoretical Formulations: I. V J. P. Zubek (ur.), *Sensory deprivation: Fifteen years of research* (str. 407–432). New York, NY: Appleton-Century-Crofts.
Zuckerman, M. (1994). *Behavioral expressions and biosocial bases of sensation seeking.* New York, NY: Cambridge University Press.
Zuckerman, M. (2005). *Psychobiology of personality* (2nd ed.). New York, NY: Cambridge University Press.
Zuckerman, M. (2007). The sensation seeking scale V (SSS-V): Still reliable and valid. *Personality and Individual Differences, 43*(5), 1303–1305.
Zuckerman, M., Eysenck, S. B. G. in Eysenck, H. J. (1978). Sensation seeking in Europe and America: Cross-cultural, age, and sex comparison. *Journal of Consulting and Clinical Psychology, 46*(1), 139–149.
Ženko, B. in Džeroski, S. (2002). Napovedovanje biorazgradljivosti z regresijskimi drevesi [Predicting biodegradability with regression trees]. *Elektrotehniški vestnik, 69*(1), 60–68.

Prispelo/Received: 29.11.2007
Sprejeto/Accepted: 28.09.2008