Výskumný ústav detskej psychológie a patopsychológie
Cyprichova 42, 831 05 Bratislava
02/4342 0973, vudpap@vudpap.sk, www.vudpap.sk

ZÁVEREČNÁ SPRÁVA
Z VÝSKUMNEJ ÚLOHY ZA ROK 2018

1.Základné informácie o výskumnej úlohe:

Názov výskumnej úlohy:
P160 - Identifikácia rizikových skupín žiakov a predikcia ich rizikového správania

Termín realizácie: 1/2017 – 12/2018
Vedúci výskumnej úlohy: Martin Kanovský/Martin Hulín
Výskumný cieľ:
Cieľom je vytvoriť predikčný model rizikového správania (špecificky skupín rizikového správania), kde ako prediktory vystupujú intelektové, osobnostné, čitateľské výkony a ďalšie socio-demografické premenné a identifikovať tie najlepšie. Predikčný model využíva moderné princípy strojového učenia, tzv. machine learning.

2. Detailný popis realizovaného výskumu:

Ide o sekundárnu analýzu dát, pričom zber dát prebehol v rámci národného projektu KOMPOSYT. Využívame údaje z digitálnej platformy Komposyt, prostredníctvom ktorej sa realizoval v roku 2015 zber údajov za účelom štandardizácie testov. K tomuto účelu bola vybraná reprezentatívna vzorka žiakov ZŠ prostredníctvom proporčného, stratifikovaného a náhodného výberu, celkovo 3364 žiakov (pri tvorbe vzorky sa vychádzalo z populačných údajov žiakov ZŠ dostupných na UIPŠ). Pre naše účely sekundárnej analýzy dát spĺňa podmienku vyplnenia všetkých relevantných testov (Eysenckov osobnostný dotazník pre deti, Test štruktúry inteligencie, Čítaciu skúšku, Škálu školského správania a škálu rizikového správania) celkovo 882 žiakov zo 7., 8. a 9. ročníka.

Spracovanie dát.

Údaje sú spracovávané v štatistickom prostredí R (R Core Team, 2017) s využitím príslušných balíkov rpart, ranger, caret, caTools, flexclust, ggplot2.

- Overenie súvislostí medzi rizikovými faktormi a problémovým a rizikovým správaním prostredníctvom korelácií, parciálnych korelácií, robustných korelácií, Bayesovskými koreláciami s využitím balíkov WRS (Wilcox & Schönbrodt, 2014), BayesFactor (Morey & Rouder, 2015), BayesMed (Nuijten, Wetzels, Matzke, Dolan & Wagenmakers, 2015).
- Modelovanie latentných tried problémového a rizikového správania a jeho Bayesovské alternatívy s využitím balíkov poLCA (Linzer & Lewis, 2011), MultiLCIRT (Bartolucci, Bacci & Gnaldi, 2016), BayesLCA (White & Murphy, 2014).
- Multivariačné regresné modely latentných tried (aj s využitím štrukturovaného modelovania) (Rosseel, 2012).
- Viac-úrovňové modely pre identifikáciu zoskupovacích efektov s využitím balíkov lm4 (Bates, Maechler, Bolker & Walker, 2015), robustlmm (Koller, 2016), nlme (Pinheiro, Bates, DebRoy, Sarkar & R Core Team, 2017).

Zistenia

Pre naše účely sekundárnej analýzy dát spĺňa podmienku vyplnenia všetkých relevantných testov (Eysenckov osobnostný dotazník pre deti, Test štruktúry inteligencie, Čítaciu skúšku, Škálu školského správania a škálu rizikového správania) celkovo 882 žiakov zo 7., 8. a 9. ročníka. Našou závislou premennou bola premenná rizikové správanie, ktorá pozostáva zo všetkých položiek dotazníka rizikového správania, pretože sa ukazuje ako dostatočne jednotná a psychometricky adekvátna na modelovanie.

Na dátach z dotazníka rizikového správania žiakov (46 položiek, 7 subškál) bola overená sila spoločného faktora (koeficientom reliability McDonaldovou omegou hierarchickou) a miera jednodimenzionálnosti (koeficientom ECV – explained common variance) tohto konštruktu využitím exploračnej bifaktorovej IRT analýzy (implementovanej v balíku TAM pre štatistický program R). Oba parametre boli dostatočne uspokojivé na to, aby sme mohli uvažovať o jednotnom a jednodimenzionálnom konštrukte rizikového správania a mohli použiť tieto dáta pre IRT analýzy a modelovanie (výsledok bifaktorovej IRT analýzy je uvedený v prílohe č.1). V prípade omegy hierarchickej spoločný faktor vysvetľoval takmer 90% rozptylu v rizikovom správaním, a zo spoločného vysvetleného rozptylu tiež takmer 90% vysvetľuje spoločný faktor. Následne sme realizovali analýzu diskrétnych latentných tried rizikového správania žiakov v škole založenej na IRT prístupe (s využitím balíka MultiLCIRT pre štatistické prostredie R). Ukázalo sa, že žiakov možno podľa vzorov odpovedí na latentnej premennej rizikového

správania rozdeliť do 4 skupín, pričom v najrizikovejšej skupine patrí asi 10% žiakov. Porovnanie modelov pre rôzny počet diskrétnych latentných tried ako aj porovnanie latentných tried z hľadiska dosiahnutého hrubého skóre rizikového správania uvádzame v prílohe č.2.

Vzhľadom na povahu zberu dát po triedach, školách a okresoch sme tiež zisťovali mieru, do akej dochádza k tzv. zoskupovacím efektom, teda, či sa skupiny žiakov s rizikovým správaním nejakým spôsobom nezoskupujú v rovnakých triedach či školách. Ukazuje sa, že k takýmto efektom dochádza, a najrizikovejší žiaci sú silne zoskupený v 8 až 10 školách (príloha č.3). Celkovo vysvetľujú zoskupovacie efekty škôl asi 15% rozptylu dát v rizikovom správaní.

V rámci analýz sme ďalej stavali na predošlých zisteniach, ktoré ukázali, že žiakov možno podľa závažnosti rozdeliť do 4 latentných skín rizikového správania. Základné charakteristiky týchto skupín sú uvedené v prílohe č.1. Následne sme pristúpili k modelovaniu príslušnosti k týmto skupinám s využitím princípov strojového učenia. Medzi prediktory boli zaradené osobnostné črty (extraverzia, neurotizmus, psychotizmus), intelektové premenné (IQ), výkony v čítacej skúške ako aj ďalšie socio-demografické premenné. Postupne sme porovnávali viaceré modely založené na algoritmoch rozhodovacích stromoch (vrátane random forest algoritmu), ordinálnej regresii, kNN algoritme a ďalších. Využili sme taktiež klastrovacie ako aj optimalizačné postupy na vyladenie parametrov modelu (tzv. cross-validácia), čo zabraňuje tzv.“overfittingu“. Ako základná porovnávacia úroveň pre overovanie presnosti týchto modelov je % respondentov v najčastejšej rizikovej skupine. V našom prípade je to 2 riziková skupina (stredne nízke riziko), kde je zhruba 39% respondentov. Naše modely dosahujú mierne vyššiu predikčnú presnosť okolo 42%. Overených bolo už desiatka modelov.

Následne sme pokračovali v rozbehnutých analýzach, pričom sme sa orientovali prevažne na aplikovanie algoritmov strojového učenia na predikciu 4 rizikových skupín žiakov, ktoré boli identifikované metódou latentných tried. V našom ponímaní predikcie vychádzame z modelov zaoberajúcich sa predikciou budúceho správania, teda zameriavame sa na čo najväčšiu presnosť predikcií na úkor pochopenia či vysvetlenia možných mechanizmov (explanácia vs, predikcia správania). Ukazuje sa totiž, že psychologické teórie majú buď veľmi nejasnú alebo iba slabú schopnosť predikovať budúce správania a zaoberajú sa primárne vysvetlením a popísaním mechanizmov správania. A hoci typicky a používajú a pomenovávajú používané regresné techniky ako predikčné, takmer nikdy sa model netestuje na nových dátach, a dochádza tak vždy k výraznému preceneniu schopnosti presne predikovať (tzv. overfitting).

Celkovo pri modelovaní s využitím strojového učenia používame cca 20 prediktorov. Našou závislou premennou sú 4 kategórie (latentné triedy) rizikového správania (nízko, stredne-nízko, stredne-vysoko, vysoko rizikoví žiaci).

Základným princípom strojového učenia, ktorí používame je, že dáta sme náhodne rozdelili na trénovacie (60% dát) a testovacie (40% dát). Následne vylaďujeme predikčný model na trénovacích dátach. Ak máme dobre vyladený model (výber iba najrelevantnejších prediktorov), overíme si presnosť jeho predikcií na testovacích dátach. Základným kritériom výberu vhodného modelu je predovšetkým predikčná presnosť. Druhé kritérium, ktoré sme použili bola čo najmenšia penalizácia modelu (využívame dva druhy penalizačných matíc, ktoré definujú, ktoré nesprávne predikcie sú penalizované a ako). Zavádzame teda aj penalizovanie nepresných predikcií, pričom niektoré predikcie sú viac penalizované ako iné.

Hlavným algoritmom strojového učenia, ktorý využívame sú tzv. klasifikačné rozhodovacie stromy, ktoré sú priamo určené na predikciu kategorických dát. Ich veľkou výhodou oproti napr. regresným technikám je, že si nevyžadujú žiadne nároky či podmienky na dáta či prediktory. Taktiež automaticky modelujú aj veľmi zložité interakcie premenných a sú veľmi ľahko interpretovateľné. Umožňujú nám navyše do algoritmu zakomponovať a zohľadniť penalizáciu určitých typov predikcií (napr. predikovanie vysokej rizikovosti hoci respondent má nízku apod.). Ako nadstavbu a elimináciu určitých obmedzení rozhodovacích stromov využíva aj algoritmus „random forest“, ktorý buduje množstvo rozhodovacích stromov, ktoré nakoniec hlasujú o výsledku predikcie. Často to vedie k výraznému zlepšeniu predikčnej presnosti, avšak nevýhodou je nemožnosť jednoducho interpretovať tieto modely. Oba tieto algoritmy ešte kombinujeme s klastrovacími technikami, kedy sa respondenti zaradia do podobných skupín a následne sa realizuje predikcia rizikových skupín na každej skupine zvlášť. Nakoniec sa výsledky predikcií za každú skupinu zlúčia a zistí sa konečná presnosť predikcie. Všetky tieto techniky sú zároveň optimalizované cross-validizáciou, aby sa eliminoval overfitting, teda príliš dobrá presnosť na trénovacích dátach ale slabá na testovacích dátach, teda zlepšenie zovšeobecniteľnosti predikčného modelu.

Následne sme overovali zhruba 55 modelov, ktoré predstavujú rôznu kombináciu prediktorov. Pre každý model boli overované predikcie vyššie spomenutých techník. Najlepšie modely dosahujú predikčnú presnosť okolo 43%, čo je oproti tzv. baseline predikčnej úrovne 39% iba mierne zlepšenie, a zatiaľ je to príliš málo na praktické využitie.

Pripravili sme revíziu a skrátenie škály rizikového správania tak, aby bola dostatočne jednodimenzionálna a praktickejšia na použitie.

Pripravila sa ďalšia publikácia do zahraničného karentovaného časopisu zameraná na objasnenie štruktúry dimenzií Eysenckovej osobnostnej teórie (extraverzia, neurotizmus,

psychotizmus, lži skóre). Článok je následne východiskom pre ďalšiu publikáciu zameranú na predikčné schopnosti osobnostných dimenzií a rizikových skupín žiakov. Pre objasnenie štruktúry jednotlivých dimenzií aj ich psychometrických kvalít boli použité bifaktorové exploračné IRT analýzy, vrátane vyhodnotenia jednodimenzionality a možnosti spočítavať hrubé skóre (miera sýtenia spoločného faktora). Predbežné výsledky z faktorových analýz sú uvedené v prílohe.

Výsledok týchto analýz zároveň prispieva do dlhoročnej diskusii o povahe extraverzie, ktorá bola často kritizovaná ako zmes dvoch nezávislých dimenzií. Prispievame k tejto diskusii použitím nových metód (bifaktorová analýza s geomin ortogonálnou rotáciou ako presnejšou metódou oproti Schmid-Leimanovej ortogonalizácii). Výsledky skutočne naznačujú na výraznú nejednotnosť extraverzie, čo má dopady na jej skórovanie a interpretáciu. Ostatné dimenzie sa ukazujú ako dostatočne psychometricky kvalitné pre použitie v praxi.

Publikovanie výskumnej úlohy :

Aktívna účasť na konferencii SAV „Psychologické dni 2017 – Agresia vo verejnom priestore“ dňa 6.-8.9.2017. Názov príspevku: Predikcia rizikového správania v školách prostredníctvom analýzy latentných tried.

Aktívna účasť na konferencii FiF UK „Osobnosť v kontexte kognícií, emócií a motivácie“ dňa 29.11.2017 v Bratislave. Názov príspevku: *Možnosti využitia princípov "machine learning" v predikcii rizikových skupín žiakov*

Príloha

-bifaktorová štruktúra psychotizmu

	G	F1	F2	F3			
P01	0,398	-0,030	0,246	0,104	0,770	χ^2	190,200
P02	0,440	-0,108	0,496	-0,014	0,548	df	116,000
P03	0,395	-0,143	-0,144	-0,234	0,748	p	0,000
P04	0,575	0,364	-0,082	0,004	0,530	CFI	0,989
P05	0,364	0,532	-0,155	-0,029	0,559	TLI	0,983
P06	0,531	0,023	-0,511	-0,131	0,439	RMSEA	0,015
P07	0,614	-0,143	0,231	-0,082	0,542	RMSEA CI	0.011/0,018
P08	0,635	-0,338	-0,007	-0,236	0,427	SRMR	0,033
P09	0,421	0,090	0,548	0,008	0,515		
P10	0,532	-0,283	0,037	-0,055	0,633		
P11	0,510	-0,027	0,532	0,028	0,456		
P12	0,418	0,415	0,196	0,091	0,606		
P13	0,530	0,107	0,000	0,487	0,470		
P14	0,598	0,042	0,132	0,358	0,496		
P15	0,463	0,157	-0,003	0,082	0,754		
P16	0,625	0,015	0,047	0,438	0,415		
P17	0,712	-0,149	-0,041	-0,188	0,434		
P18	0,515	-0,140	-0,089	0,147	0,686		
P19	0,364	0,365	-0,287	0,043	0,651		
P20	0,359	0,066	-0,026	0,209	0,822		
	9,999	0,815	1,120	1,030	11,501		
	99,98	0,664225	1,2544	1,0609	102,9595		
mean	0,500	0,041	0,056	0,052	114,461		
				omega	0,89952		
				omegaH	0,873489		
				ECV	0,61237		

-bifaktorová štruktúra extraverzie

E01	0,254	0,051	0,310	-0,066	0,832	χ^2	335,600
E02	0,315	0,159	0,012	0,149	0,853	df	132,000
E03	0,281	-0,121	0,350	0,130	0,767	p	0,000
E04	0,311	0,772	-0,028	-0,034	0,306	CFI	0,980
E05	0,345	-0,046	0,193	0,311	0,745	TLI	0,969
E06	0,165	0,325	-0,041	0,305	0,772	RMSEA	0,023
E07	0,369	0,292	0,003	0,034	0,777	RMSEA CI	0,020/0,026
E08	0,311	-0,040	0,812	0,041	0,241	SRMR	0,031
E09	0,491	0,434	-0,034	-0,008	0,569		
E10	0,340	-0,018	0,094	0,112	0,862		
E11	0,299	-0,107	0,859	0,005	0,161		
E12	0,255	-0,150	0,093	0,525	0,628		
E13	0,695	0,269	-0,035	-0,109	0,432		
E14	0,737	0,058	-0,021	-0,122	0,438		
E15	0,218	0,736	-0,082	-0,119	0,390		
E16	0,182	-0,132	0,170	0,364	0,788		
E17	0,255	-0,106	-0,013	0,645	0,507		
E18	0,532	0,365	-0,008	-0,148	0,562		
E19	0,553	0,108	0,031	0,219	0,634		
E20	0,625	0,085	-0,101	-0,047	0,590		
E21	0,554	-0,041	-0,002	0,110	0,679		
	8,087	2,893	2,562	2,297	12,533		
	65,39957	8,369449	6,563844	5,276209	85,60907		
mean	0,385	0,138	0,122	0,109	98,142		
				omega	0,872297		
				omegaH	0,666376		
				ECV	0,435633		

-bifaktorová štruktúra subdimenzie impulzivity (subdimenzia extraverzie)

E01	0,304	0,518	0,202	-0,026	0,598	χ^2	10,130
E03	0,581	-0,034	-0,002	-0,043	0,660	df	6,000
E05	0,605	-0,029	-0,241	0,078	0,569	p	0,119
E08	0,626	0,058	0,573	-0,026	0,276	CFI	0,999
E10	0,305	0,163	-0,043	0,115	0,865	TLI	0,995
E11	0,662	-0,008	0,671	-0,028	0,111	RMSEA	0,015
						RMSEA	
E12	0,328	-0,058	-0,006	0,562	0,573	CI	0/0,031
E16	0,357	-0,222	0,003	0,251	0,761	SRMR	0,010
E17	0,277	0,011	-0,103	0,652	0,488		
	4,045	0,399	1,054	1,535	4,901		
	16,36203	0,159201	1,110916	2,356225	19,98837		
mean	0,449444	0,044333	0,117111	0,170556	24,889		
				omega	0,803089		
				omegaH	0,65739		
				ECV	0,495199		

-bifaktorová štruktúra subdimenzie sociability (subdimenzia extraverzie)

E02	0,492	0,081	0,060	0,111	0,748	χ^2	17,190
E04	0,629	0,487	0,063	-0,064	0,359	df	17,000
E06	0,366	0,044	-0,048	-0,257	0,796	p	0,441
E07	0,499	0,058	-0,142	-0,026	0,727	CFI	1,000
E09	0,659	0,068	0,620	-0,035	0,175	TLI	1,000
E13	0,707	-0,022	-0,023	0,252	0,435	RMSEA	0,002
						RMSEA	
E14	0,627	-0,172	-0,107	0,417	0,392	CI	0/0,017
E15	0,554	0,621	0,002	-0,022	0,307	SRMR	0,013
E18	0,600	0,130	0,084	0,190	0,580		
E19	0,563	-0,265	0,106	-0,019	0,602		
E20	0,527	-0,083	0,006	0,385	0,567		
E21	0,457	-0,298	-0,034	0,125	0,686		
	6,680	0,649	0,587	1,057	6,374		
	44,6224	0,421201	0,344569	1,117249	46,50542		
mean	0,556667	0,054083	0,048917	0,088083	52,879		
				omega	0,879462		
				omegaH	0,843852		
				ECV	0,677136		

-bifaktorová analýza dimenzie neurotizmus

N01	0,533	0,036	-0,186	-0,056	0,677	χ^2	335,83
N02	0,383	-0,067	0,242	0,304	0,698	df	132
N03	0,491	0,274	-0,165	0,025	0,656	p	0
N04	0,378	0,147	-0,071	0,091	0,822	CFI	0,984
N05	0,473	0,012	0,045	0,114	0,761	TLI	0,975
N06	0,407	0,357	-0,012	0,135	0,688	RMSEA	0,023
						RMSEA	
N07	0,664	0,150	0,385	-0,022	0,389	CI	0,020/0,026
N08	0,577	-0,118	-0,038	0,323	0,547	SRMR	0,030
N09	0,599	-0,162	0,031	0,044	0,612		
N10	0,385	0,582	-0,036	-0,017	0,512		
N11	0,411	-0,133	-0,016	-0,033	0,812		
N12	0,368	0,264	0,183	0,057	0,758		
N13	0,431	0,012	-0,008	0,397	0,656		
N14	0,562	0,503	0,358	-0,045	0,301		
N15	0,535	-0,002	0,211	-0,007	0,669		
N16	0,230	0,001	0,035	0,035	0,945		
N17	0,753	-0,036	-0,184	-0,137	0,379		
N18	0,612	-0,075	0,072	-0,186	0,580		
N19	0,622	-0,112	0,054	-0,087	0,590		
N20	0,403	-0,313	-0,048	0,323	0,633		
N21	0,515	0,141	-0,109	0,032	0,702		
	10,332	1,461	0,743	1,290	13,387		
	106,7502	2,134521	0,552049	1,6641	111,1009		
mean	0,492	0,070	0,035	0,061	124,488		
				omega	0,892463		
				omegH	0,857515		
				ECV	0,707484		

- Bifaktorová analýza dimenzie lži skóre

L01	0,678	0,175	-0,328	-0,004	0,403	χ^2	125,780
L02	0,728	0,233	-0,105	0,117	0,390	df	62,000
L03	0,487	0,339	0,025	0,045	0,646	p	0,000
L04	0,459	0,306	0,143	0,117	0,662	CFI	0,996
L05	0,283	-0,344	0,277	0,057	0,722	TLI	0,992
L06	0,593	0,067	-0,019	0,211	0,599	RMSEA	0,018
						RMSEA	
L07	0,620	-0,043	0,374	-0,094	0,465	CI	0,014/0,023
L08	0,584	0,400	0,023	-0,070	0,493	SRMR	0,021
L09	0,510	0,046	0,442	-0,197	0,504		
L10	0,566	-0,034	0,124	-0,022	0,663		
L11	0,574	-0,014	0,008	-0,260	0,603		
L12	0,645	-0,155	-0,273	-0,009	0,486		
L13	0,734	-0,145	-0,004	-0,092	0,432		
L14	0,648	-0,262	0,066	0,172	0,478		
L15	0,667	-0,214	-0,112	-0,071	0,492		
L16	0,718	-0,010	-0,069	0,407	0,314		
	9,494	0,345	0,572	0,307	8,352		
	90,13604	0,119025	0,327184	0,094249	90,67649		
mean	0,593	0,022	0,036	0,019	99,028		
				omega	0,915664		
				omegaH	0,910207		
				ECV	0,763188		