Vysoké učení technické v Brně
Fakulta informačních technologií

Projekt do predmetu ZZN

Získavanie znalostí nad vybranou vzorkou dát

Firc Anton (xfirca00)
Lapšanský Tomáš (xlapsa00)

1 Zadanie

Táto sekcia obsahuje popis dát určených k dolovaniu a formuláciu úlohy pre dolovanie nad popísanými dátami.

Popis dát

Názov datasetu: *Databáze z výzkumu konzumace alkoholu*

Dataset obsahuje dáta o študentoch z dvoch Portugalských stredných škôl *Gabriel Pereira secondary school* a *Secondary School Mousinho da Silveira*, a to dáta o rodine, ich výsledkoch v škole a hlavne konzumácii alkoholu. Dáta pozostávajú z dvoch súborov a to:

- *student-mat.csv* - študenti matematiky
- *student-por.csv* - študenti portugalčiny

Oba súbory obsahujú rovnaké dáta:

- základné údaje (pohlavie, vek, adresa)
- rodinné údaje
  - veľkosť rodiny
  - rodinný stav rodičov
  - dosiahnuté vzdelanie matky a otca
  - povolanie matky a otca
  - študentov právny zástupca (opatrovník)
  - vzťahy v rodine
- údaje o škole
  - škola ktorú študent navštevuje
  - dôvod výberu navštevovanej školy
  - dĺžka cesty do školy
  - týždenný počet hodín strávených štúdiom
  - počet predchádzajúcich neúspešných ukončení predmetu
  - doučovanie sa (v škole, vrámci rodiny, platené)
  - mimoškolské aktivity
  - predchádzajúce a ďalšie plánované vzdelanie
  - počet absencií
  - známky z predmetu (prvý polrok, druhý polrok, výsledná známka)
- prístup k internetu
- partnerské vzťahy
- množstvo voľného času popri štúdiu
- množstvo času stráveného s priateľmi

- konzumácia alkoholu cez týždeň / víkend
- zdravotný stav

V oboch datasetoch sa nachádza niekoľko rovnakých študentov, títo študenti sú identifikovateľný podľa identických atribútov ktoré charakterizujú každého študenta.

Úloha

Vyššie spomínané stredné školy sa rozhodli spustiť program na prevenciu konzumácie alkoholu u ich študentov. Zaujíma ich teda, aké aspekty života (rodina, škola, zázemie, okolie, ...) najviac ovplyvňujú množstvo konzumovaného alkoholu u študentov. Na základe týchto výsledkov chcú študentom pomôcť prispôsobiť podmienky tak aby minimalizovali konzumáciu alkoholu. Zároveň chcú dokázať predikovať ktorí zo všetkých študentov by mohli byť náchylný k vysokej miere konzumácie alkoholu a poskytnúť im väčšiu pozornosť.

Z tejto špecifikácie požiadavkov vyplývajú dve pod-úlohy pre dolovanie dát:

- zistiť ktoré atribúty majú najväčší vplyv na mieru konzumácie alkoholu u študenta
- vytvoriť model pre predikciu miery konzumácie alkoholu

2 Riešenie

Táto sekcia obsahuje popis riešení jednotlivých úloh spolu s krokmi ktoré boli vykonané pre získanie týchto výsledkov.

2.1 Úloha 1

Pre riešenie prvej úlohy sme zvolili korelačnú analýzu, kde sme zisťovali ktoré atribúty majú najväčší vplyv na mieru konzumácie alkoholu. Všetky atribúty sme rozdelili do štyroch skupín podľa oblastí do ktorých patria:

- rodina
  - *famsize* - veľkosť rodiny
  - *pstatus* - rodinný stav rodičov
  - *medu* - dosiahnuté vzdelanie matky
  - *fedu* - dosiahnuté vzdelanie otca
  - *mjob* - pracovné odvetvie matky
  - *fjob* - pracovné odvetvie otca
  - *guardian* - právny zástupca študenta
  - *famsup* - podpora pri vzdelávaní od rodiny
  - *famrel* - vzťahy v rodine
- zázemie
  - *address* - miesto bydliska

  - *traveltime* - dĺžka cesty do školy
  - *paid* - extra platené hodiny v rámci predmetu
  - *nursery* - navštevoval materskú školu
  - *internet* - internetová prípojka doma
- škola
  - *school* - zvolená škola
  - *studytime* - čas strávený štúdiom
  - *failures* - počet neúspešných zakončení predmetu
  - *schoolsup* - extra podpora pri vzdelávaní
  - *activities* - mimoškolské aktivity
  - *absences* - počet vymeškaných hodín
  - *higher* - plánuje vyššie vzdelanie
  - *reason* - dôvod zvolenia aktuálnej školy
- osobné
  - *sex* - pohlavie
  - *age* - vek
  - *romantic* - má vzťah
  - *freetime* - voľný čas po škole
  - *goout* - ako často chodí von s priateľmi

Pred akoukoľvek analýzou dát, bolo potrebné dáta najprv pripraviť tak aby boli vhodné pre dolovanie. Keďže sa dáta nachádzali v dvoch samostatných datasetoch, bolo najprv potrebné spojiť tieto datasety do jedného. Spojenie datasetov sme vykonali použitím kombinácie operátorov *Join*, *Filter Examples* a *Append* tak aby došlo k správnemu spojeniu oboch datasetov, a nechýbali žiadne hodnoty. Pred spojejím datasetov, bolo ešte potreba premenovať niektoré atribúty ktoré mali v oboch datsetoch rovnaké názvy, ale niesli inú informáciu, a to konkrétne *failures, absences, G1, G2, G3*. Keďže nie každý študent mal zapísané oba predmety a v datasete vzniklo veľa chýbajúcich hodnôt práve pri atribútoch odpovedajúcim predmetom, rozhodli sme sa tieto atribúty zlúčiť a vytvorili priemery pre známky a absencie. Aby sa nestratila informácia o tom, aké predmety mal študent zapísané, vytvorili sme nový polynomiálny atribút udávajúci ktoré predmety má študent zapísané.

Najprv, sme zisťovali do akej miery ovplyvňuje konzumáciu alkoholu zapísaný predmet, prípadne škola. Zistili sme, že študenti navštevujúci školu *Secondary School Mousinho da Silveira* môžu byť takisto náchylnejší k vyššej konzumácii alkoholu. Čo sa týka predmetu, miera konzumácie alkoholu závisí primárne od školy. Rozdiely sú však minimálne, ako ukazuje tabuľka 1 a ukazuje sa to aj pri korelačnej analýze, kde škola alebo zapísaný predmet nefigurujú ako dôležité atribúty.

	Obe školy		GP		MS	
	Davg	Wavg	Davg	Wavg	Davg	Davg
Portugalština	1.531	2.255	1.528	2.213	1.532	2.274
Matematika	1.440	2.240	1.167	2	2.143	2.857
Oba predmety	1.476	2.291	1.450	2.272	1.700	2.450

Tabuľka 1: Priemerná miera konzumácie alkoholu podľa predmetu a školy.

Zisťovanie závislosti medzi mierou konzumácie alkoholu a ostatnými atribútmi sme zisťovali pomocou korelačnej analýzy. Aby bolo možné vykonať korelačnú analýzu bolo nutné previesť polynomiálne atribúty na numerické. Vybrali sme pre každú mieru konzumácie alkoholu atribúty ktoré majú najväčší vplyv, a následne sme tieto atribúty rozdelili do kategórií podľa oblasti ktorej sa dotýkajú.

Týždňová miera konzumácie alkoholu

Najväčší vplyv na mieru konzumácie alkoholu má pohlavie, čas strávený štúdiom, to či študent plánuje vyššie vzdelanie, množstvo voľného času, to koľko chodí von s priateľmi, absencie a známky. Najviac atribútov spadá do oblasti *osobné*, tesne nasleduje *rodina*. Presná korelácia je zaznačená v nasledujúcej tabuľke:

sex (M)	0.282		osobné
sex (F)		- 0.282	osobné
Medu (2)		- 0.095	rodina
Fjob (services)	0.093		rodina
reason (other)	0.119		škola
reason (reputation)		- 0.099	škola
guardian (other)	0.091		rodina
higher (yes)		- 0.121	škola
higher (no)	0,121		škola
freetime (3)		- 0.101	osobné
freetine (5)	0.112		osobné
goout (2)		- 0.114	osobné
goout(5)	0.207		osobné
absences_avg	0.138		škola
grades_avg		- 0.153	škola

Tabuľka 2: Atribúty ktoré zvyšujú a znižujú týždennú mieru konzumácie alkoholu v najvyššej miere a oblasť do ktorej tento atribút spadá.

Víkendová miera konzumácie alkoholu

Na víkendovú mieru konzumácie alkoholu má znova veľký vplyv pohlavie, oproti týždennej sa zvyšuje vplyv toho ako veľmi chodí študent von s priateľmi a pridáva sa jeho zdravotný stav. Najviac atribútov spadá znovu do oblastí *osobné* a *rodina*. Presná korelácia je zaznačená v nasledujúcej tabuľke:

sex (M)	0.315		osobné
sex (F)		- 0.315	osobné
famsize (LE3)	0.094		rodina
famsize (GT3)		- 0.094	rodina
Fedu (0)		- 0.101	rodina
Fjob (services)	0.106		rodina
studytime (1)	0.212		škola
studytime (3)		- 0.212	škola
goout (2)		- 0.214	osobné
goout(5)	0.309		osobné
health (5)	0.099		osobné
health (1)		- 0.087	osobné
absences_avg	0.126		škola
grades_avg		- 0.147	škola

Tabuľka 3: Atribúty ktoré zvyšujú a znižujú víkendovú mieru konzumácie alkoholu v najvyššej miere a oblasť do ktorej tento atribút spadá.

Záver

Najväčší vplyv vrámci oboch mier konzumácie alkoholu má pohlavie, chlapci sú náchylnejší k vyššej miere konzumácie alkoholu. Takisto s veľkým rozdielom od ostatných atribútov je aj miera toho ako veľmi chodí študent von s priateľmi.

Atribúty spadajúce do oblasti *rodina* ukazujú, že na mieru konzumácie alkoholu môže mať vplyv dosiahnuté vzdelanie matky alebo otca, prípadne otcove zamestnanie. Takisto sa ukazuje súvislosť veľkosti rodiny s mierou konzumácie alkoholu.

Atribúty z oblasti *škola* ovplyvňujúce mieru konzumácie alkoholu sú čas strávený štúdiom, známky a absencie. Rolu môže hrať aj to, či študent plánuje pokračovať vo vyššom vzdelaní. U atribútov známky a absencie je podľa nás vhodné polemizovať nad tým do akej miery ovplyvňujú mieru konzumácie alkoholu, alebo sú mierou konzumácie alkoholu ovplyvnené. V každom prípade však môžu slúžiť ako vhodný atribút pre predikciu miery konzumácie alkoholu.

Atribúty z kategórie *rodina* nemajú výrazný vplyv na mieru konzumácie alkoholu. Dobré rodinné vzťahy môžu pozitívne ovplyvniť mieru konzumácie alkoholu. Zaujímavým zistením je, že pracovné odvetvie otca môže ovplyvniť mieru konzumácie alkoholu.

Nakoniec ešte zisťujeme aký vplyv na seba majú týždenná a víkendová miera konzumácie alkoholu. Ukázalo sa, že víkendová a týždenná miera konzumácie alkoholu majú na seba navzájom v porovnaní s ostatnými atribútmi veľký vplyv a to presne 0.611.

Pre obe kategórie teda ešte stanovuje spoločné atribúty ktoré najviac vplývajú na mieru konzumácie alkoholu: *sex, goout, absences_avg, grades_avg.*

Podľa zistených výsledkov sa ukazuje ako prínosné zamerať sa na oblasť života týkajúcu sa školy, nakoľko rodinné atribúty ako vzdelanie rodičov a ich zamestnanie je podľa našich predpokladov mimo vplyv školy. Pokiaľ škola dokáže motivovať študentov tak aby sa rozhodli pokračovať vo vyššom vzdelaní a trávili viac času štúdiom a menej času chodením von s priateľmi mala by klesnúť aj miera konzumácie alkoholu. Pokiaľ by tieto opatrenia nemohli byť zavedené pre všetkých študentov, vidíme ako vhodné začať s chlapcami a na škole *Secondary School Mousinho da Silveira* nakoľko sa ukázalo, že ich miera konzumácie alkoholu Študentov s takýmito atribútmi býva vyššia.

2.2 Úloha 2

Druhá úloha má za cieľ predikovať mieru konzumácie alkoholu u študentov. Miera konzumácie alkoholu je daná triedami $1 - 5$, takže sa jedná o problém *klasifikácie*. O konzumácii alkoholu máme dve informácie, týždňová a víkendová miera konzumácie alkoholu. Z týchto hodnôt sme sa rozhodli vypočítať celkovú mieru konzumácie alkoholu, nakoľko ako ukázali výsledky prvej úlohy, týždňová a víkendová miera konzumácie alkoholu od seba v prevažnej miere závisia a zároveň nevidíme žiadnu pridanú informačnú hodnotu pri rozlíšení týchto mier pre splnenie požiadaviek úlohy. Pre výpočet celkovej miery konzumácie alkoholu sme zvolili vzorec :

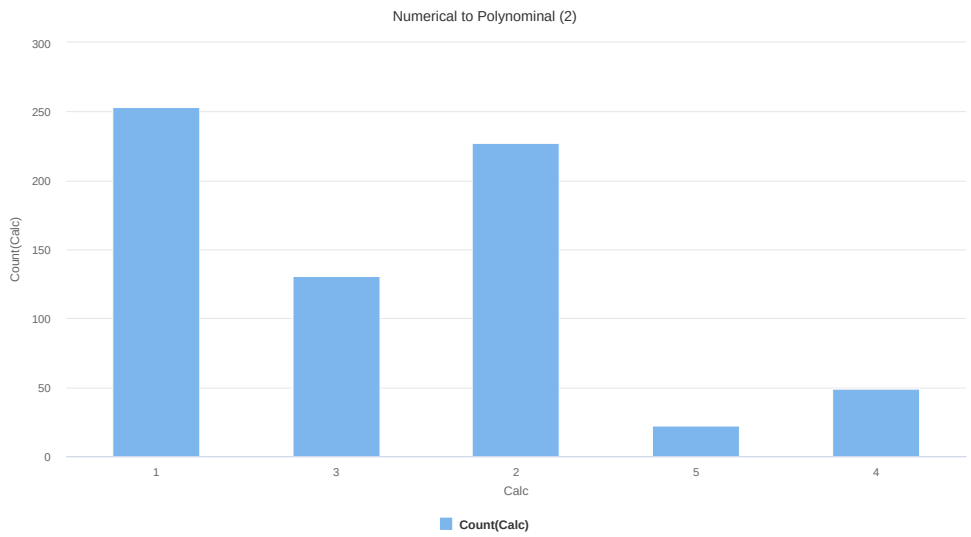
$$Calc = ceil(avg(Dalc, Walc))$$

Kontrola pomocou korelačnej tabuľky ukázala, že celková miera konzumácie alkoholu je závislá o trochu viac od víkendovej konzumácie alkoholu (8.809 vs 0.928). Tento fakt je spôsobený výpočtom, kedy zaraďujeme celkovú mieru konzumácie alkoholu do jednej z už existujúcich tried zaokrúhlením nahor, a tým, že víkendová miera konzumácie býva u študentov obvykle vyššia. Ďalej skúmame, do akej miery sa líši závislosť celkovej miery konzumácie alkoholu od závislostí týždennej a víkendovej miery. Podľa výsledkov uvedených v tabuľke 4 zisťujeme, že sa atribúty vplývajúce na celkovú mieru konzumácie alkoholu nijak nelíšia od atribútov vpývajúcich na týždennú alebo víkendovú mieru, preto tento atribút budeme predikovať.

sex (M)	0.31		osobné
sex (F)		- 0.31	osobné
famsize (LE3)	0.089		rodina
famsize (GT3)		- 0.089	rodina
Medu (2)		- 0.11	rodina
Fedu (0)		- 0.101	rodina
Fjob (services)	0.117		rodina
reason (other)	0.095		škola
studytime (1)	0.215		škola
studytime (3)		- 0.147	škola
higher (yes)		- 0.094	škola
higher (no)	0.094		škola
freetime (5)	0.087		osobné
freetime (3)		- 0.093	osobné
goout (2)		- 0.186	osobné
goout(5)	0.28		osobné
health (5)	0.094		osobné
health (1)		- 0.075	osobné
absences_avg	0.148		škola
grades_avg		- 0.149	škola

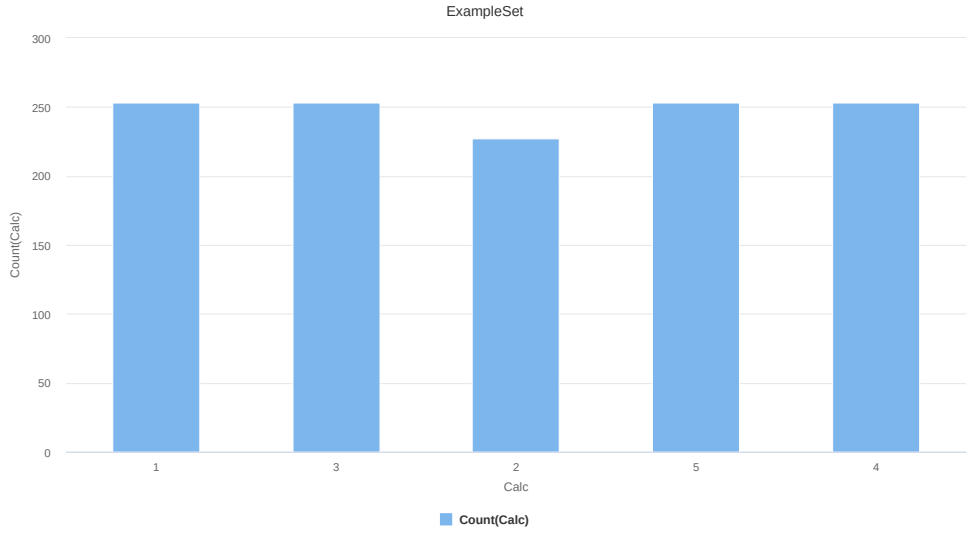
Tabuľka 4: Atribúty ktoré zvyšujú a znižujú celkovú mieru konzumácie alkoholu v najvyššej miere a oblasť do ktorej tento atribút spadá.

Príprava dát pre predikciu prebehla obdobne ako pri korelačnej analýze. Jediný krok ktorý nebol pri predikcii vykonaný je prevod polynomiálnych atribútov na numerické. Po príprave dát, tak aby neobsahovali žiadne chýbajúce hodnoty, sme analyzovali rozloženie do jednotlivých tried. Zistili sme, že je dataset veľmi nevyvážený, ako ukazuje obrázok 1.



Obr. 1: Zastúpenie jednotlivých tried

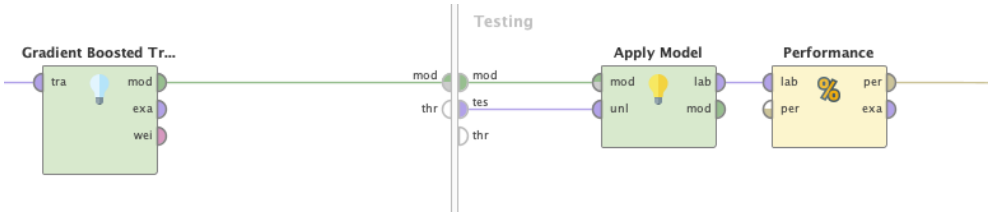
Pre získanie správnych výsledkov je dataset najprv potreba vybalancovať. Prvou možnosťou, je použitie operátora *Sample*, tento operátor však iba oreže majoritné triedy (vykoná *downsampling*) a v datasete následne zostalo z takmer 700 vzoriek, menej než 100. Ďalšou možnosťou, je použitie operátora *Generate Weights* ktorý vzorkám priradí váhu. Takto označené vzorky však nie sú vhodné pre všetky prediktívne modely. Poslednou možnosťou, je *upsampling* minoritnej triedy, pre tieto účely bol použitý operátor *SMOTE upsampling* z rozšírenia Operator Toolbox. Správnym použitím tohoto operátora sa nám podarilo vybalancovať dataset, a neredukovať vzorky. Zastúpenie jednotlivých tried po vybalancovaní je znázornené v obrázku 2.



Obr. 2: Zastúpenie jednotlivých tried po vybalancovaní

Po vybalancovaní dát mohli byť dáta predané jednotlivým modelom pre tréning.

Trénovanie a vyhodnotenie modelu sme zapúzdrili do bloku *Cross Validation*. Tento blok zabezpečuje rozdelenie dát na trénovacie a testovacie, tréning a následné vyhodnotenie modelu. Vnútri tohoto bloku sa následne nachádza použitý model, operátor *Apply Model* pre použitie tohoto modelu nad testovacími dátami a operátor *Performance* ktorý vyhodnotí výkon modelu. Architektúra je znázornená na obrázku 3. Optimalizáciu parametrov modelu resp. atribútov sme realizovali pomocou operátorov *Optimize parameters* resp. *Optimize selection*.



Obr. 3: Architektúra bloku *Cross Validation*

2.2.1 Gradient Boosted Tree

Experimentovaním s rôznymi modelmi sme ako prvý model zvolili *Gradient Boosted Tree*. Skúšaním sa ukázalo nastavenie parametrov pre dosiahnutie najlepších výsledkov, ako ukazuje obrázok 4. Tieto nastavenia potvrdilo aj hľadanie najlepších parametrov modelu pomocou optimalizácie evolučnými algoritmami. Takto nastavený model dosiahol presnosť 71.30% ako ukazuje obrázok 5. Ďalej sme sa pokúsili optimalizovať vstupné atribúty pre dosiahnutie lepšieho výsledku. Experimentovaním, ani použitím evolučných algoritmov pre optimalizáciu výberu atribútov sme však nedosiahli lepšie skóre.

Obr. 4: Použité parametre modelu *Gradient Boosted Tree*

accuracy: 71.30% +/- 5.19% (micro average: 71.31%)

	true 1	true 3	true 2	true 5	true 4	class precision
pred. 1	134	25	92	0	4	52.55%
pred. 3	24	179	34	0	4	74.27%
pred. 2	73	27	76	4	8	40.43%
pred. 5	5	3	7	234	0	93.98%
pred. 4	7	9	11	5	227	87.64%
class recall	55.14%	73.66%	34.55%	96.30%	93.42%	

Obr. 5: Presnosť klasifikácie modelom *Gradient Boosted Tree* s optimalizáciou parametrov a bez optimalizácie atribútov.

2.2.2 Random Forest

Model *Random Forest* dosiahol po experimentoch s parametrami modelu a bez optimalizácie atribútov presnosť 71.56%. Parametre modelu sme nastavili podľa výsledkov experimentu na : *počet stromov* : 300, *maximálna hĺbka* : 15, ostatné parametre zostali bez zmeny. Následne sme vyskúšali optimalizáciu parametrov pomocou evolučných algoritmov, dosiahli sme rovnaké výsledky ako pri našich experimentoch. Posledne sme skúsili optimalizovať atribúty znovu použitím evolučných algoritmov, dosiahli sme mierne zlepšenie a to 72.88% ako ukazuje obrázok .

accuracy: 71.56% +/- 4.53% (micro average: 71.56%)

	true 1	true 3	true 2	true 5	true 4	class precision
pred. 1	142	26	112	0	3	50.18%
pred. 3	19	180	30	0	1	78.26%
pred. 2	70	14	54	1	2	38.30%
pred. 5	6	8	9	241	1	90.94%
pred. 4	6	15	15	1	236	86.45%
class recall	58.44%	74.07%	24.55%	99.18%	97.12%	

Obr. 6: Presnosť klasifikácie modelom *Random Forest*.

accuracy: 72.88% +/- 3.65% (micro average: 72.88%)

	true 1	true 3	true 2	true 5	true 4	class precision
pred. 1	146	20	106	0	6	52.52%
pred. 3	25	192	31	1	1	76.80%
pred. 2	70	22	73	0	5	42.94%
pred. 5	3	10	8	252	1	91.97%
pred. 4	9	9	9	0	240	89.89%
class recall	57.71%	75.89%	32.16%	99.60%	94.86%	

Obr. 7: Presnosť klasifikácie modelom *Random Forest* s optimalizáciou atribútov.

2.2.3 Deep Learning

Použitie hlbokej neurónovej siete s optimalizáciou atribútov dosiahlo presnosť 64% ako ukazuje obrázok 8. Ako aktivačná funkcia bola zvolená funkcia *Maxout* ktorá v experimentoch dosahovala najlepšie výsledky, zároveň sme nastavili počet epoch algoritmu na 50. Optimalizáciou parametrov modelu alebo atribútov sa nám však nepodarilo dosiahnuť lepší výsledok.

accuracy: 64.00% +/- 4.79% (micro average: 64.00%)

	true 1	true 3	true 2	true 5	true 4	class precision
pred. 1	121	42	90	3	7	46.01%
pred. 3	30	128	30	7	7	63.37%
pred. 2	89	54	86	2	8	35.98%
pred. 5	3	15	3	237	10	88.43%
pred. 4	10	14	18	4	221	82.77%
class recall	47.83%	50.59%	37.89%	93.68%	87.35%	

Obr. 8: Presnosť klasifikácie modelom *Deep Learning*.

2.2.4 Neurónová sieť

Vyskúšali sme použiť pre predikciu aj neurónovú sieť. Keďže neurónová sieť nevie pracovať s polynomiálnymi atribútmi, bolo potreba previesť všetky atribúty na číselné. S takto upraveným datasetom sme dosiahli presnosť 65.38% ako ukazuje obrázok 9. Optimalizáciou atribútov alebo parametrov modelu sa nám však nepodarilo dosiahnuť lepší výsledok.

accuracy: 65.38% +/- 4.42% (micro average: 65.38%)

	true 1	true 3	true 2	true 5	true 4	class precision
pred. 1	121	27	75	7	5	51.49%
pred. 3	37	157	40	10	14	60.85%
pred. 2	82	38	87	4	12	39.01%
pred. 5	3	11	7	229	6	89.45%
pred. 4	10	20	18	3	216	80.90%
class recall	47.83%	62.06%	38.33%	90.51%	85.38%	

Obr. 9: Presnosť klasifikácie modelom *Neural Network* bez optimalizácie parametrov modelu a atribútov.

3 Záver

V rámci projektu sme vyriešili dve úlohy za použitia prostredia RapidMiner. Prvá úloha spočívala v zistení atribútov ktoré najviac ovplyvňujú mieru konzumácie alkoholu. Korelačnou analýzou sme zistili, že najväčší vplyv na mieru konzumácie alkoholu majú atribúty *sex, goout, absences_ avg, grades_ avg*. Podľa týchto zistení sme navrhli kroky ktoré môžu školy podniknúť pre v snahe znížiť mieru konzumácie alkoholu u svojich študentov.

Cieľom druhej úlohy bolo predikovať mieru konzumácie alkoholu u študentov. Najlepší výsledok 72.88% sme dosiahli použitím modelu *Random Forest* a optimalizáciou jeho atribútov evolučnými algoritmami. Kvalita výsledkov je do veľkej miery ovplyvnená nevyváženosťou datasetu a s tým súvisiace malé množstvo dát. V pôvodnom, neupravenom, datasete je minoritná trieda študentov najviac konzumujúcich alkohol veľmi malá, má cca 20 vzoriek. Ostatné triedy sú tiež nevyvážené, preto bolo potrebné generovať vzorky dokopy z troch tried a to vo veľkom objeme (180 - 80 podľa triedy). Predpokladáme, že tento fakt má najväčší vplyv na kvalitu výsledkov.