MENDELOVA UNIVERZITA V BRNĚ

Klasifikácia spotreby alkoholu medzi študentami strednej školy

Provozně ekonomická fakulta

Bc. Lenka Bistáková
Bc. Diana Brnovik

6. 6. 2020

Obsah

1. Téma

1.1. Druh problému

Alkohol každoročne ovplyvňuje milióny študentov stredných či vysokých škôl. Univerzitné roky sú jedným z najpopulárnejších období experimentovania s alkoholom. Mnoho mladých dospelých pripúšťa pitie alkoholu ešte pred vstupom na vysokú školu. Po ukončení strednej školy a odchode z domu chcú zažiť svoju novoobjavenú slobodu a nezávislosť. Dostupnosť alkoholu pri športových podujatiach a spoločenských aktivitách je pre študentov často lákavá.

Štúdia skúmala spotrebu alkoholu medzi študentami kurzov matematiky a španielčiny na strednej škole v spojení s rôznymi sociálnymi, rodovými či študijnými údajmi. Prostredníctvom výberu z množstva atribútov bližšie popisujúcich vzťah medzi množstvom skonzumovaného alkoholu a napríklad vzťahu s rodinnými príslušníkmi je možné použiť tieto dáta pre klasifikáciu a následnú predikciu závažnosti požívania alkoholu medzi študentami stredných škôl.

1.2. Cieľ analýzy dát

Cieľom tohto projektu je vytvorenie neurónovej siete určenej pre klasifikáciu. Ide o klasifikačný problém o 4 triedach, vzostupne definujúcich závažnosť požívania alkoholu u študentov. K riešeniu tohto problému bude použitý programovací jazyk Python a knižnice pre vytváranie neurónových sietí a knižnice pre strojové učenie ako *Neurolab* a *Scikit Learn*.

Pomocou troch rôznych riešení neurónovej siete budú vytvorené tri prototypy a v závislosti na použitých knižniciach rôzne učiace algoritmy. Tie budú testované pre rôzne hladiny lambda, počty neurónov a pomeru rozdelenia dát na trénovacie, validačné a testovacie. Následne budú tieto výsledky prehľadne vizualizované a bude zhodnotené, ktorý prototyp prináša najlepšie výsledky.

2. Dáta a nástroje

2.1. Zdroje dát

V projekte pracujeme s dátami získanými zo štúdie na stredných školách dostupných na portáli `https://data.world/data-society/student-alcohol-consumption`. Dáta obsahujú dva súbory, pričom jeden obsahuje 395 záznamov o študentoch matematiky a druhý 649 záznamov o študentoch portugalčiny. Pre ďalšiu prácu s *datasetom* sme použili súbor vo formáte CSV so 649 záznamami.

Na základe štúdiou získaných dát boli vlastnosti klasifikujúce výslednú konzumáciu zoradené do 33 stĺpcov a obsahujú informácie o:

- type školy a dôvodu výberu,
- pohlaví, veku a adrese študenta,
- informáciách o rodinnom stave a rodinných príslušníkoch,
- vzťahu s rodinnými príslušníkmi a romantickom vzťahu,
- podmienkach pre štúdium – napr. prístup k internetu, trvanie cesty do školy, podpora štúdia vo forme doučovaní a pod.,
- úspechoch študenta v podobe známok, absencií a počtu neprejdených predmetov,
- denná a víkendová konzumácia alkoholu.

Ako klasifikačný atribút sme zvolili atribúty *Dalc* a *Walc* špecifikujúce množstvo požitého alkoholu v týždni a počas víkendov. Miera požívania je určená číslami od 1–5, 1 značiaca veľmi malú mieru, 5 značiaca veľmi vysokú mieru alkoholu. Všetky dáta *datasetu* sú vyjadrené numericky a jednotlivé významy značenia sa nachádzajú v prílohách práce.

2.2. Použité nástroje

Pre implementáciu neurónovej siete sme sa rozhodli použiť programovací jazyk *Python* a populárnu distribúciu tohoto *Data Scienca* jazyka – *Anaconda*. Umožňuje vytvorenie virtuálneho prostredia pre vývoj a zjednodušuje správu balíkov. V súvislosti s ňou bolo nutné využiť jednu z jej prostredí – *Jupyter Notebook*. Predstavuje webovú aplikáciu a zároveň interaktívne kódovacie prostredie, obohacujúc text o paragrafy, obrázky a iné prvky RTF.

Pre prvý prototyp neurónovej siete bola využitá knižnica *Neurolab* so základnými algoritmami pre neurónové siete s flexibilnou konfiguráciou a učebnými algoritmami. Druhý prototyp bol zostavený pomocou knižnice pre strojové učenie *Scikit learn*, vybavenú rôznymi algoritmami klasifikácie či regresie. Triedy *MLPClassifier* (klasifikátor pre *multi-layer perceptron* siete) a *LinearRegression* boli použité pre druhý a tretí prototyp. *Scikit* spolupracuje s knižnicou *Numpy* podporujúcu prácu s multidimenzionálnymi vektormi a maticami. Pre vizualizácie dát bola využitá knižnica *Matplotlib*, ktorá sfunkčňuje prácu s grafmi ako v programe *Matlab*. Okrem nich sme využili aj ostatné knižnice ako napr. *Seaborn* či *Pandas*.

2.3. Práca s dátami

Rozloženie dát a preprocessing

Prvým krokom k získaniu dát, na ktorých sa môže neurónová sieť učiť, bolo odstránenie redundantných a nepotrebných dát. Rozhodli sme sa z pôvodných 33 stĺpcov odstrániť 9, ktoré vzhľadom k rozloženiu dát boli pre učenie siete nevypovedajúce. Jednalo sa napríklad o vlastnosti ako navštevovaná materská škola, trvanie cesty zo školy, semestrálne známky (výsledná známka bola ponechaná), adresa, či dôvod výberu strednej školy. Klasifikačné atribúty boli spojené do jedného atribútu s názvom *alcohol_consumption* vzniknutého priemerom týždennej a víkendovej konzumácie, pričom víkendová konzumácia má vyššiu váhu:

$$\frac{2 * (Walc + 1) * Dalc}{3}$$

Čo sa týka rozloženia dát, dáta boli podľa tried 1–5 nerovnomerne rozložené a pre 4 a 5 triedu pripadalo menej ako 100 záznamov – jednalo sa teda o tzv. *skewed data*. Z tohto dôvodu sme sa rozhodli tieto triedy spojiť do jednej triedy s označením 4 (vysoká a veľmi vysoká miera komzumácie). Zo 649 záznamov, 245 patrí do 1 triedy, 216 do druhej triedy, 121 do tretej triedy a do 4 spadá 71 záznamov. Výsledný CSV súbor tak obsahuje 24 atribútov, bližšie špecifikovaných v prílohách práce.

Aby neurónová sieť dokázala pracovať optimálne, vyžaduje na vstupe normalizované dáta. To si vyžaduje normalizáciu dát pomocou škálovania, ktoré transformuje dáta tak, že sa hodnoty nachádzajú v rozsahu [0,1]. K tejto transformácii bola použitá funkcia *MinMaxScaler()* knižnice *Sklearn*.

```
### Pre-process data using MinMaxScaler() [0,1]

min_max_scaler = preprocessing.MinMaxScaler()
X_scaled = min_max_scaler.fit_transform(X)
```

Zdrojový kód 1: Pre-process dát

Pre prácu neurónovej siete s triedami bolo potrebné, aby boli hodnoty klasifikačných tried premapované na jednoznačne určené hodnoty [0,0], [0,1], [1,0] a [1,1] nasledovne:

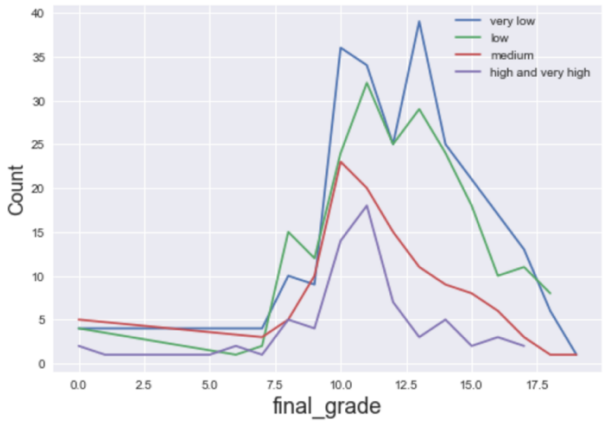
```
### Target transform

# class 1--4 --> alcohol consumption (1=low, 4=high)
val_map = {1: [0,0], 2: [0,1], 3: [1,0], 4: [1,1]}
T = np.array([val_map[y] for y in Y])
data['alcohol_consumption'] = data['alcohol_consumption'].map( {1: [0,0], 2: [0,1], 3: [1,0], 4: [1,1]})
```

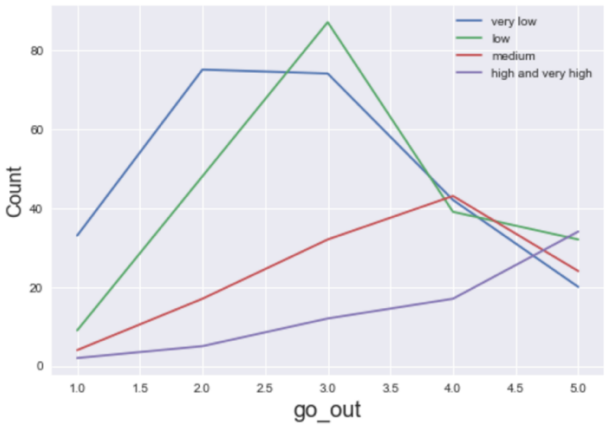
Zdrojový kód: 2: Target transformácia

Vizualizácia závislosti dát

Pre znázornenie závislostí medzi jednotlivými vlastnosťami boli využité dva typy grafov. Prvým bolo znázornenie rozloženia a zároveň závislosti tried konzumácie alkoholu na jednotlivých atribútoch prostredníctvom 4 trendov predstavujúcich jednotlivé triedy. Na obr. 1 môžme vidieť závislosť finálnej známky študenta od počtu výskytov danej triedy v *datasete*. Môžme pozorovať, že triedy 1 a 2 (nízke miery konzumácie) sú početnejšie a majú veľmi podobný trend, zatiaľ čo triedy 3 a 4 (vyššia miera konzumácie) sú menej početné. Existuje tendencia, že študenti s vyššou konzumáciou majú známky nižšie. Podobne je to aj s vlastnosťou popisujúcou frekvenciu chodením von, kde je možné pozorovať zvyšujúcu sa tendenciu chodiť von s pribúdajúcou konzumáciou alkoholu.

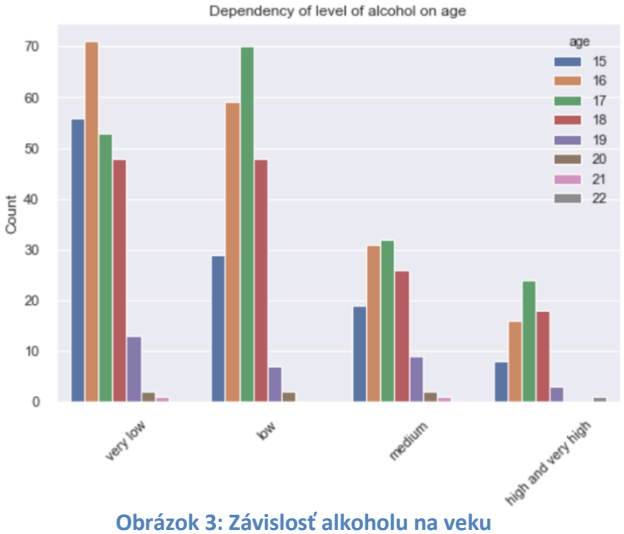


Obrázok 1: Známky

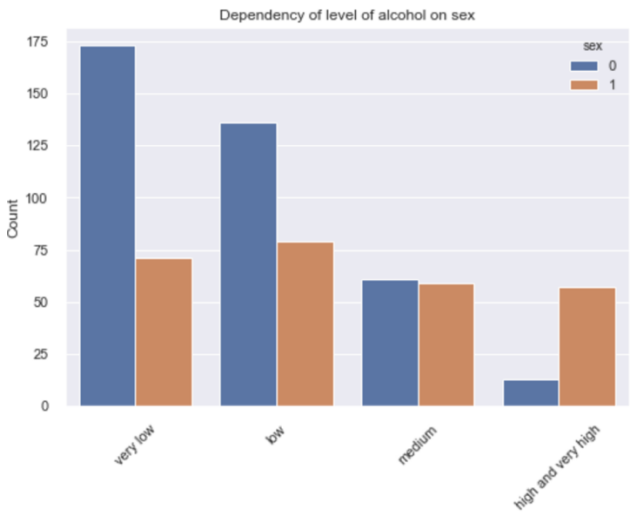


Obrázok 2: Frekvencia chodenia von

Druhé zobrazenie je v podobe histogramov. Na nasledujúcich obrázkov sú znázornené závislosti miery konzumácie od veku a pohlavia. Z grafu môžeme usúdiť, že tieto dva atribúty budú značnejšie vplývať na výsledok neurónovej siete, nakoľko napriek rôznemu rozloženiu dát v *datasete* sa pre každú triedu hodnoty menia na rozdiel ostatných histogramov. Z prvého grafu možno vidieť okrem väčšej početnosti vekov 16 a 17 prevahu 16-ročných študentov, ktorí konzumujú alkohol menej a 17-ročných, ktorí prevažujú v ostatných triedach. Ďalej môžeme pozorovať, že tendencia zvýšeného alkoholu prevažuje viac u mužov ako u žien.



Obrázok 3: Závislosť alkoholu na veku



Obrázok 4: Závislosť alkoholu na pohlaví

Rozdelenie dát

Pre zistenie výkonnosti neurónovej siete podľa pomerov dát, na ktorých sa bude učiť a na ktorých bude svoje učenie overovať, sme sa rozhodli dáta rozdeliť vo viacerých pomeroch, ktoré boli postupne aplikované pre všetky tri modely. Rozloženia dát na trénovacie, validačné a testovacie je nasledovné:

- 70% trénovacích, 30% testovacích – zo 70% tvorí 20% validačných dát, t.j. pomer (56/14/30) [%]
- 80% trénovacích, 20% testovacích – z 80% tvorí 20% validačných dát, t.j. pomer (64/16/20) [%]
- 90% trénovacích, 10% testovacích – z 90% tvorí 20% validačných dát, t.j. pomer (72/18/10) [%]

Tieto pomery museli byť delené dvoma spôsobmi kvôli odlišnému prístupu knižníc k deleniu dát. Pre prvý prototyp *(Neurolab)* bolo delenie uskutočnené funkciou `np.split()`. Neurónová sieť sa učila na 415 záznamoch (príp. 364 al. 467), simulovanie prebiehalo na 104 záznamoch (príp. 91 al. 116) a predikovala na 130 záznamoch (príp. 194 al. 66).

```
split_sizes = [int(len(X_process)*train_fraction), int(len(X_process)*(train_fraction+validation_fraction))]
X_train,X_validation,X_test = np.split(X_scaled, split_sizes)
T_train,T_validation,T_test = np.split(T, split_sizes)
```

Zdrojový kód 3: Split dát na 3 časti

Nakoľko knižnica *Sklearn* zabezpečovala automatické rozdelenie trénovacích dát na trénovacie a validačné, postupovali sme s rozdelením iným spôsobom a dáta sme rozdelili len v pomeroch 80/20 (resp. 70/30 al. 90/10)

```
X2_train, X2_test, T2_train, T2_test = train_test_split(
    X_process, T,
    train_size=train_fraction+validation_fraction,
    test_size=test_fraction
)
```

Zdrojový kód 4: Split dát na 2 časti

3. Implementácia

3.1. Učiace algoritmy

Vzhľadom k tomu, že každá z použitých knižníc umožňovala iné nastavenie parametrov, v rámci implementácie učiacich algoritmov sme sa rozhodli použiť dve rôzne knižnice.

Knižnica *Neurolab* umožňuje vytvoriť neónovú sieť doprednú neurónovú sieť (angl. *feedforward).* Parametrami sú dáta (angl. *features*), vrstvy siete, resp. ich počty neurónov. Ďalej je možné nastaviť funkcie jednotlivých vrstiev. V našom prípade bola použitá jedna skrytá vrstva, ktorej počet neurónov vychádzal z matematického vzorca. Funkciu skrytej aj východzej vrstvy neurónovej siete sme nastavili na logaritmickú sigmoidálnu transferovú funkciu. Následne sú parametrami trénovania tejto neurónovej siete tréningové dáta (angl. *train features*), tréningové ciele (angl. *train targets*), epochy, miera učenia (angl. *learning rate*), cieľ tréningu a nastavenia výpisu výstupu. Neurónová sieť nemá možnosť nastavenia validačných dát. Po natrénovaní neurónovej siete je možné simulovať klasifikáciu testovaných dát. Z tohto dôvodu sme doimplementovali vlastnú validáciu, viď v prílohách. Vlastná validácia riadi učenie neurónovú sieť v iteráciách. Jedna iterácia obsahuje 10 epoch. Následne spustí simuláciu klasifikácie na validačnej množine dát a jej úspešnosť porovná s úspešnosťou predchádzajúcej iterácie cyklu. Iterácie pokračujú dovtedy pokým nie je dosiahnuté stanovené maximum počtu iterácií alebo pokým rozdiel medzi aktuálnou a predchádzajúcou úspešnosťou presahuje nami stanovenú toleranciu úspešnosti. Výsledkom každého trénovania neurónovej siete sú hodnoty funkcie chybovosti SSE, tj. súčet chýb štvorcov (angl. *error sum of squares, SSE*) voči epochám tréningu. Nevýhodou funkcie SSE je to, že táto hodnota narastá v prípade narastajúceho množstva záznamov aj keď proporcionálna chybovosť neurónovej siete nenarastá. Taktiež táto knižnica neumožňovala zobrazenie učiacich kriviek neurónovej siete vo vzťahu k rôznemu množstvu dát a teda prívetivé trénovanie na rôznom počte dát, prípadne kros-validáciu.

Ďalšou použitou knižnicou bola knižnica *Sklearn*, ktorá umožňuje vytvoriť viacvrstvový perceptrón (angl. *multilayer perceptron, MLP*), estimátor lineárnej regresie najmenších štvorcov (angl. *least squares linear regression estimator*) a iné. Táto knižnica umožňuje vykreslenie učiacich kriviek na základe predanej množiny rôznych množstiev použitých dát k trénovaniu a validácií. Učiace krivky tu zobrazujú pomer funkcie chybovosti MSE, t.j. strednú štvorcovú chybu (angl. *mean squared error, MSE*) voči množstvám trénovaných dát. Výhodou funkcie MSE je to, že chyby štvorcov sú delené počtom sčítavaných štvorcov, a preto vzrastajúce množstvo dát nesprávne nezvyšuje chybovosť tak ako v prípade SSE.

Parametrami viacvrstvového perceptrónu sú:

- aktivačná funkcia,
- alfa (L2, regularizačný parameter ~ lambda),
- počet skrytých vrstiev a ich neurónov,
- miera učenia (learning rate),
- maximálny počet iterácií,
- zamiešanie dát,
- tolerancia,
- pomer validačných dát,
- a iné.

Parametre estimátoru lineárnej regresie sú základné (napr. zrýchlenie výpočtu) a boli v tejto práci ponechané v ich predvolených hodnotách.

Nastavenie parametrov v rámci tejto práce

Pomer trénovacích, validačných a testovacích dát

Tak ako bolo spomenuté vo vyšších kapitolách pomer trénovacích, validačných dát a testovacích je pre všetky učiace algoritmy v rôznych meraniach:

- 70/30 (56/14/30) [%],
- 80/20 (64/16/20) [%]
- 90/10 (72/18/10) [%].

Počet neurónov skrytej a výstupnej vrstvy

Počet neurónov skrytej vrstvy je 8 na základe matematického vzorca (Stack Exchange Inc., 2020):

$$N_h = \frac{N_s}{(\alpha * (N_i + N_o))}$$

Ni – počet neurónov vstupnej vrstvy
No – počet neurónov výstupnej vrstvy
Ns – počet záznamov dát
α – škálovateľný faktor (zvyčajne v hodnote 2–10)

Výstupná vrstva má 2 neuróny na základe počtu cieľov predikcií (angl. *targetov*).

Počet epóch na iteráciu

Počet epóch na iteráciu doprednej neurónovej siete je 10.

Miera učenia (angl. *learning rate*)

Miera učenia je pre všetky učiace algoritmy 0.001.

Alfa (L2, regularizačný parameter)

Regularizačný parameter pre viacvrstvový perceptrón a estimátor lineárnej regresie má v meraniach rôzne hodnoty:

- 0,0001,
- 0,0005,
- 0,001,
- 0,005,
- 0,01.

Cieľ a tolerancia

Cieľ trénovania doprednej neurónovej siete je 1e-5.
Tolerancia rozdielu úspešnosti učenia medzi jednotlivými iteráciami je pre všetky algoritmy.

Maximálny počet iterácií

Maximálny počet iterácií pre všetky učiace algoritmy je 10 000 iterácií.

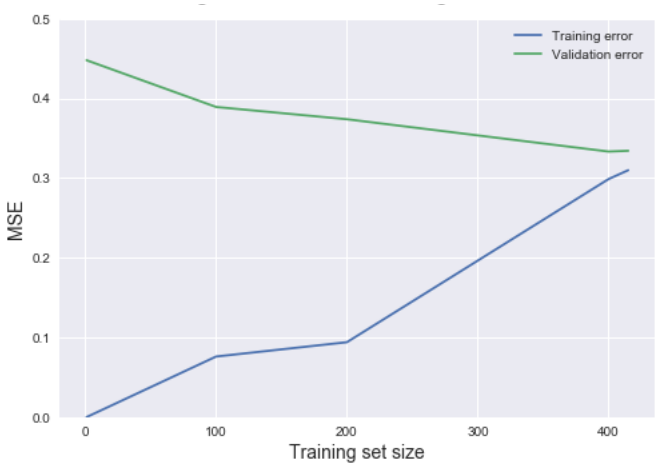
K-kros validácia

Úroveň k-kros validácie pre viacvrstvový perceptrón a estimátor lineárnej regresie je 5.

3.2. Merania

Alfa hodnoty

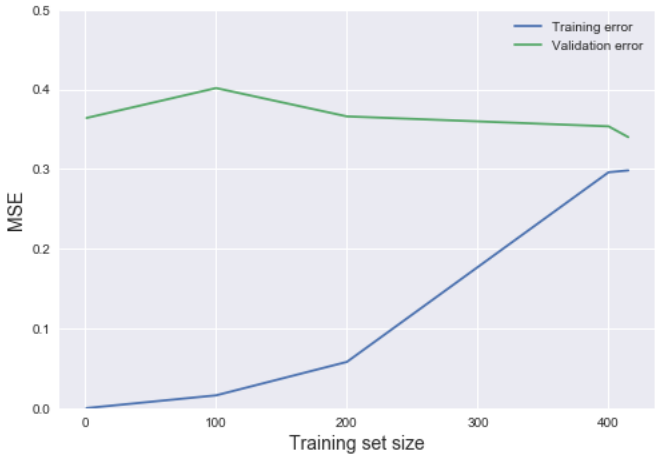
Viacvrstvový perceptrón (angl. *multilayer perceptron, MLP*) knižnice *Sklearn* umožňuje zmenu nastavenia regularizačného parametra alfa, ktorého východzia hodnota je 0,0001. (Scikit-learn developers [BSD License], 2020)
V rámci tohto projektu boli uskutočnené viaceré merania s rôznymi hodnotami alfa pri zachovaní nami zvoleného východzieho pomeru trénovacích, validačných a testovacích dát v pomere 80/20. Taktiež v týchto meraniach boli zachované všetky nami vybrané vlastnosti.



Obrázok 5: Učiaca krivka pre alfa 0,0001

33	2	0	4
37	9	0	4
13	8	1	6
6	3	0	4

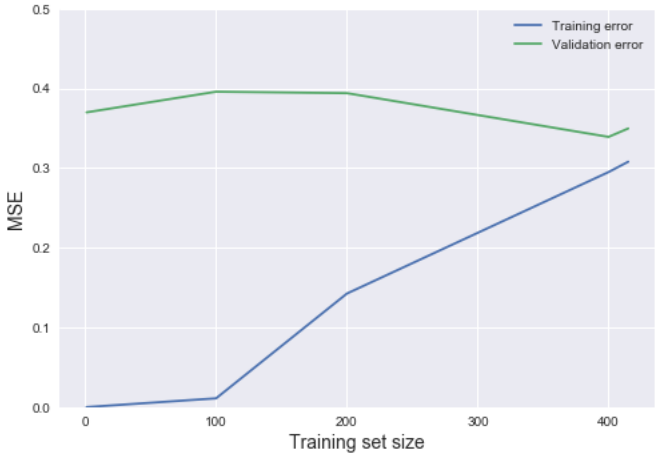
Tabuľka 1: Confusion matrix pre alfa 0,0001



Obrázok 6: Učiaca krivka pre alfa 0,0005

32	3	0	4
36	10	2	2
11	10	2	5
5	5	0	3

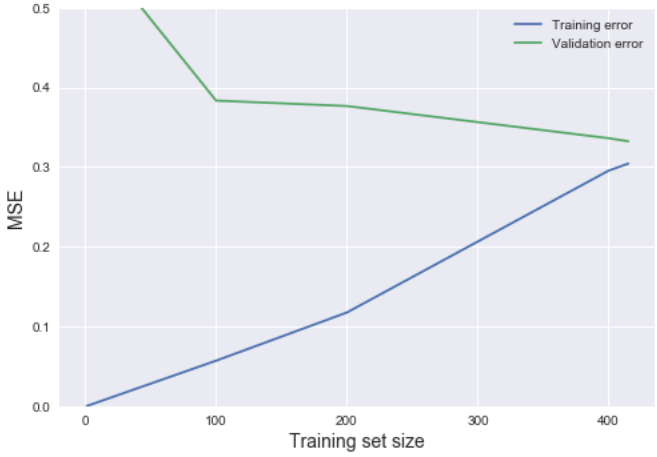
Tabuľka 2: Confusion matrix pre alfa 0,0005



Obrázok 7: Učiaca krivka pre alfa 0,001

34	1	2	2
41	6	2	1
13	6	3	6
6	3	0	4

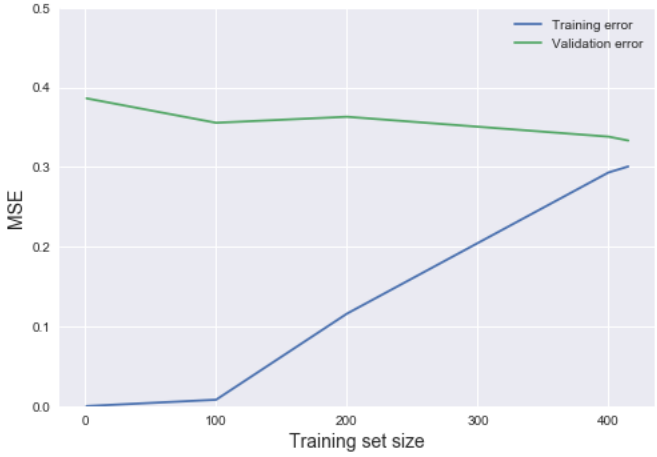
Tabuľka 3: Confusion matrix pre alfa 0,001



Obrázok 8: Učiaca krivka pre alfa 0,005

34	1	0	4
37	8	0	5
14	3	0	11
5	3	0	5

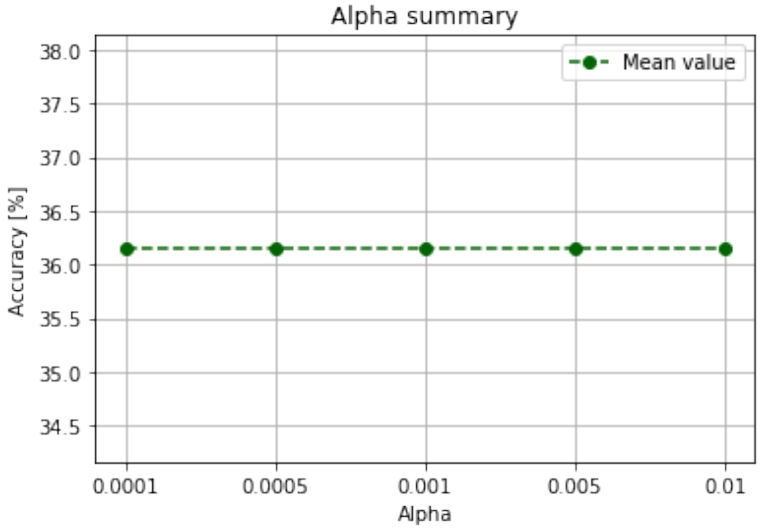
Tabuľka 4: Confusion matrix pre alfa 0,005



Obrázok 9: Učiaca krivka pre alfa 0,01

35	0	0	4
37	8	0	5
16	5	0	7
6	3	0	4

Tabuľka 5: Confusion matrix pre alfa 0,01



Obrázok 10: Porovnanie úspešnosti pri rôznych alfa

Hodnoty alfa sú naprieč meraniami jednotné, avšak samotná chybovosť je pri každom hodnotení iná a prerozdelená do iných tried. Pre ďalšie merania sme sa rozhodli vybrať hodnotu alfa 0,001 vzhľadom na približne rovnomerné počty v jej confusion matrix („nízke výkyvy“). Ak by v niektorom meraní bola hodnota úspešnosti vyššia, zvolili by sme túto hodnotu.

Pridávanie/odoberanie atribútov

Pre všetky učiace algoritmy boli následne urobené analýzy chybovosti vzhľadom na validačné testovacie dáta. Z týchto dát boli vyfiltrované vzorky s nesprávnou predikciou výsledku (angl. *target*). Následne pre každú vlastnosť bol nájdený modus v rámci týchto vyfiltrovaných dát a bol vypočítaný jeho percentuálny podiel. Vlastnosti, ktorých podiel modusu prekročil 70 % vo validačných alebo testovacích dátach, boli odstránené.

	mode	count	percentage	remove
0	0.0	65.0	0.738636	True
1	16.0	36.0	0.409091	False
2	0.0	63.0	0.715909	True
3	1.0	78.0	0.886364	True
4	1.0	35.0	0.397727	False
5	1.0	33.0	0.375000	False
6	2.0	45.0	0.511364	False
7	2.0	55.0	0.625000	False
8	0.0	53.0	0.602273	False
9	1.0	37.0	0.420455	False
10	0.0	67.0	0.761364	True
11	0.0	82.0	0.931818	True
12	1.0	60.0	0.681818	False
13	0.0	53.0	0.602273	False
14	1.0	70.0	0.795455	True
15	1.0	49.0	0.556818	False
16	0.0	53.0	0.602273	False
17	4.0	40.0	0.454545	False
18	3.0	31.0	0.352273	False
19	3.0	25.0	0.284091	False
20	5.0	26.0	0.295455	False
21	0.0	40.0	0.454545	False
22	11.0	18.0	0.204545	False

Obrázok 11: Modusy validačných dát

	mode	count	percentage	remove
0	0.0	45.0	0.542169	False
1	17.0	30.0	0.361446	False
2	0.0	55.0	0.662651	False
3	1.0	70.0	0.843373	True
4	1.0	32.0	0.385542	False
5	1.0	29.0	0.349398	False
6	2.0	27.0	0.325301	False
7	2.0	41.0	0.493976	False
8	0.0	59.0	0.710843	True
9	2.0	38.0	0.457831	False
10	0.0	62.0	0.746988	True
11	0.0	81.0	0.975904	True
12	1.0	46.0	0.554217	False
13	0.0	45.0	0.542169	False
14	1.0	66.0	0.795181	True
15	1.0	56.0	0.674699	False
16	0.0	44.0	0.530120	False
17	4.0	40.0	0.481928	False
18	3.0	24.0	0.289157	False
19	3.0	31.0	0.373494	False
20	5.0	37.0	0.445783	False
21	0.0	28.0	0.337349	False
22	10.0	18.0	0.216867	False

Obrázok 12: Modusy testovacích dát

Odstránených bolo spolu 7 vlastností (angl. *features*) a to:

- pohlavie,
- počet rodinných príslušníkov
- status vzťahu rodičov,
- zákonný zástupca,
- počtu nezvládnutých predmetov
- podpora v podobe školských doučovaní
- snaha dosiahnuť vyššie vzdelanie.

Pomery trénovacích, validačných a testovacích dát

Nasledujúce podkapitoly popisujú úspešnosť jednotlivých neurónových sietí pre rôzne pomery rozdelenia dát. Pre najúspešnejšie rozdelenie je neurónová sieť bližšie špecifikovaná prostredníctvom informáciách o úspešnosti, chybovosti, znázornenia grafov úspešnosti či závislosti chybovosti na počte epoch.

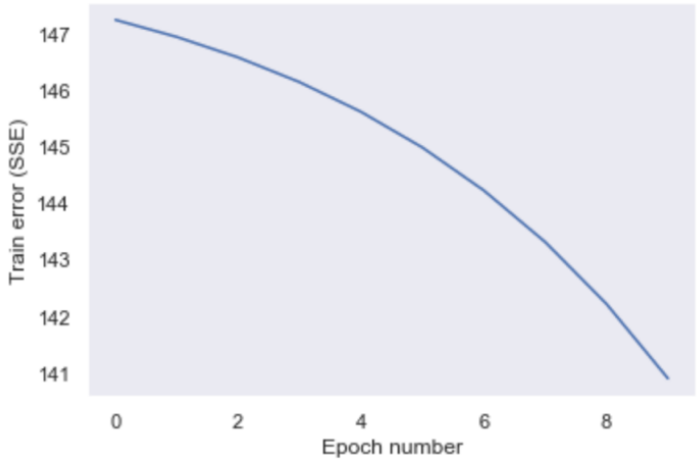
1. neurónová sieť – dopredná neurónová sieť

Výsledky merania úspešnosti prvej neurónovej siete implementovanej knižnicou *Neurolab* pre jednotlivé pomery dát môžeme vidieť v nasledujúcej tabuľke. Neurónová sieť dosahuje najlepší výsledok validácie pri pomere 80/20, a naopak pri predikovaní na základe testovacích dát dosahuje prekvapivo najhorších výsledkov. Odchýlky v úspešnosti nám tak neumožňujú jednoznačne povedať, či sa so zvyšujúcimi trénovacími dátami zvyšuje aj úspešnosť neurónovej siete. Skreslené výsledky môžu byť spojené aj s nízkym počtom záznamov *datasetu* a veľkým množstvom atribútov v pomere k počtu záznamov.

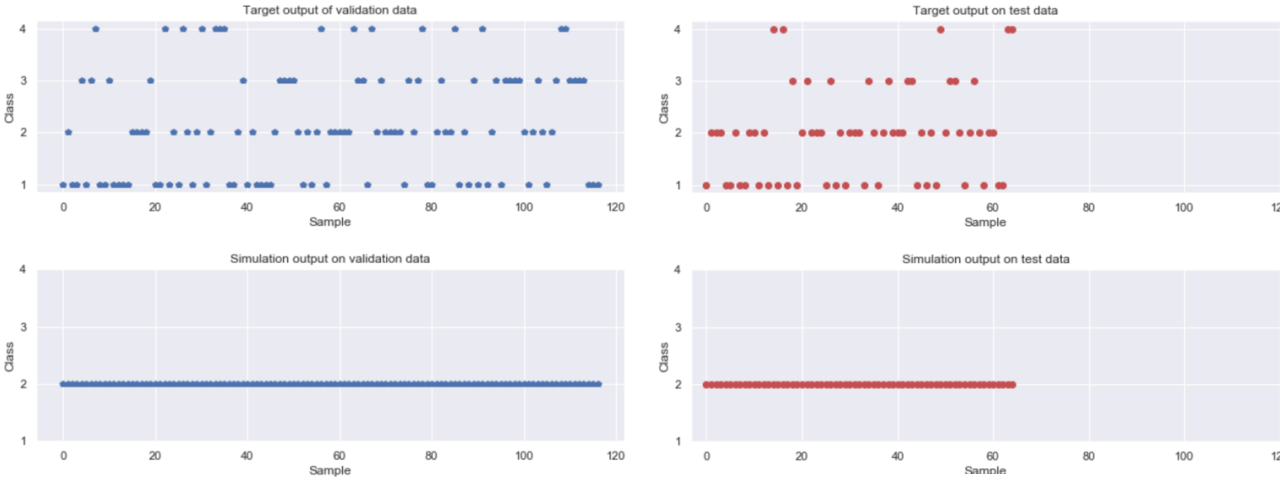
Pomery	70/30	80/20	90/10
Úspešnosť validácie	57.69 %	59.13%	53.42%
Úspešnosť testovania	61.03 %	58.46 %	63.85 %

Tabuľka 6: Pomery trénovacích, validačných a testovacích dát

K bližšej analýze tejto neurónovej siete budeme ďalej počítať s najúspešnejším pomerom na testovacích dátach a teda pomere 90/10. Neurónová sieť pri daných vstupných parametroch dosiahla koniec trénovania v druhej iterácii pri počte 10tich epoch s hodnotou chybovosti validácie 46,58%. Výsledkom tohto trénovania je chybovosť SSE s hodnotou 140,89. Znižujúcu sa chybovosť v závislosti na počte epoch znázorňuje nasledujúci graf:



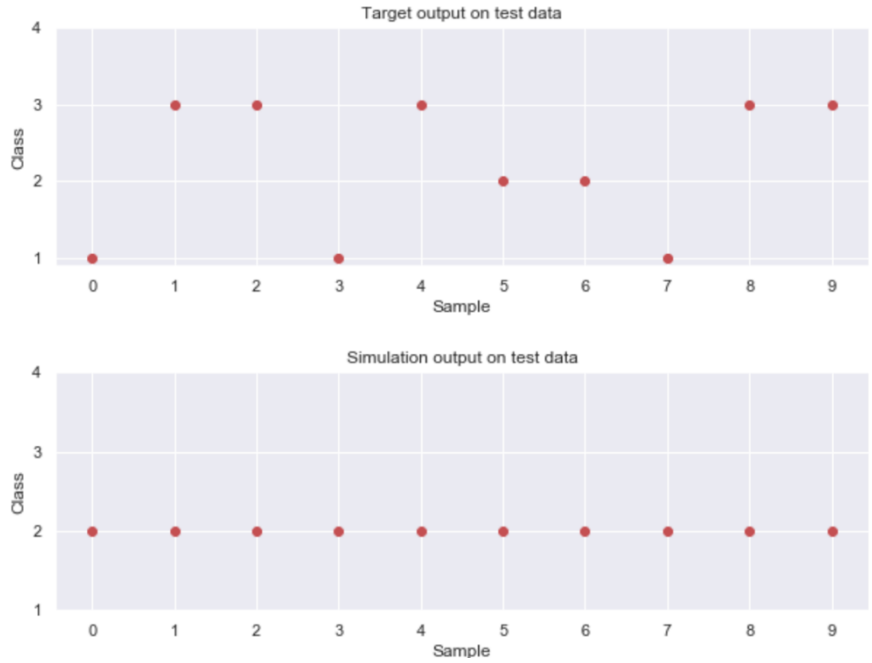
Obrázok 13: Graf závislosti chybovosti od počtu epoch



Obrázok 14: Porovnanie klasifikovaných a ich správnych hodnôt tried

Hodnoty váh a *bias* jednotlivých vrstiev sú dostupné v objekte *net.layers*. Táto knižnica však neumožňuje zobrazenie učiacich kriviek v závislosti na množstve dát. Nasledujúce vizualizácie zobrazujú pôvodné validačné a testovacie výsledky, ku ktorým by mala byť schopná neurónová sieť mieriť pri vysokej úspešnosti. Nakoľko však táto neurónová sieť nedosahuje vysokej úspešnosti, môžeme vidieť, že pri validácii aj predikcii klasifikovala výsledky ako triedu č. 2.

V neposlednom rade boli tieto vizualizácie vytvorené aj pre náhodných 10 prvkov testovacích dát. Nakoľko boli všetky záznamy opäť klasifikované ako 2 trieda, nepozorujeme podobný trend klasifikácie.



Obrázok 15: Porovnanie desiatich náhodných klasifikovaných a ich správnych hodnôt tried

2. neurónová sieť – viacvrstvový perceptrón

Druhá neurónová sieť, viacvrstvový perceptron od *Sklearn*, bola v rámci predikcie výsledkov u testovaných dát najúspešnejšia, keď pomer trénovacích a validačných dát (spolu) obsahoval 70 % pôvodných dát a pomer testovacích dát bol 30 %, viď tabuľka nižšie.

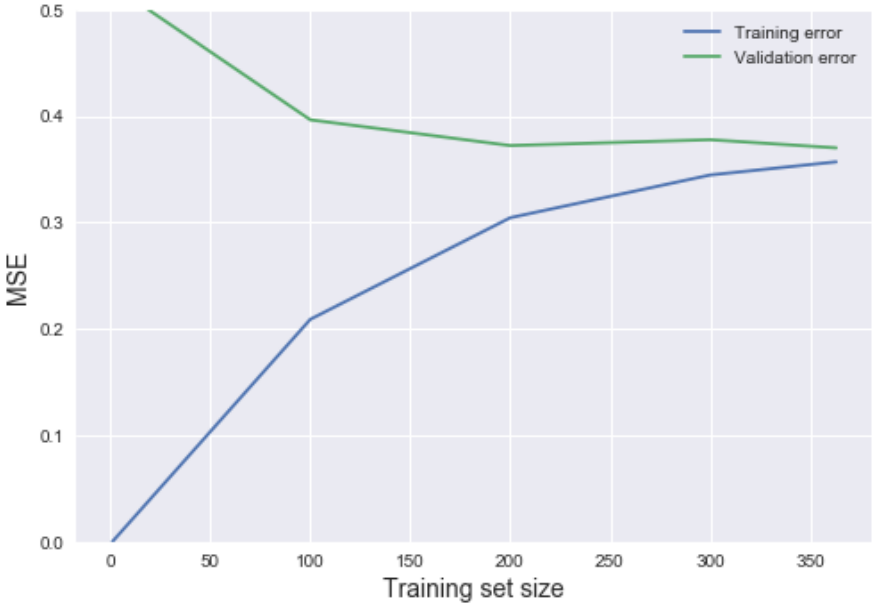
Pomery	70/30	80/20	90/10
Úspešnosť testovania	39.49 %	35.38 %	32.31 %

Tabuľka 7: Pomery trénovacích, validačných a testovacích dát

Vysoká chybovosť validačných dát (obrázok č. 16) indikuje problém s hodnotou *biasu* funkcie. Ak by bola nízka chybovosť tréningových dát, jednalo by sa o nízku hodnotu *biasu*. Avšak v tomto prípade sa jedná skôr o vysokú hodnotu *biasu*. Vysoká hodnota *biasu* naznačuje podučenie neurónovej siete.

Čo sa týka *variancie*, ak by prevažoval veľký rozdiel medzi krivkami trénovacích a validačných dát, tak by sa jednalo o nízku *varianciu*. V opačnom prípade, ak by tento rozdiel bol prevažne vysoký, tak by sa jednalo o vysokú *varianciu*. V tomto prípade sa rozdiel približne konštantne znižuje, čo nenaznačuje preučenie.

Taktiež môžeme vidieť mierny pokles úspešnosti u krivky učenia nad trénovacími dátami neurónovej siete za 300 záznamami, avšak krivka učenia nad validačnými dátami má naďalej mierne zlepšenie, a preto pri väčšom množstve dát by mohlo nastať celkové zlepšenie klasifikácie.



Obrázok 16: Učiaca krivka pre pomer dát 70/30 [%]

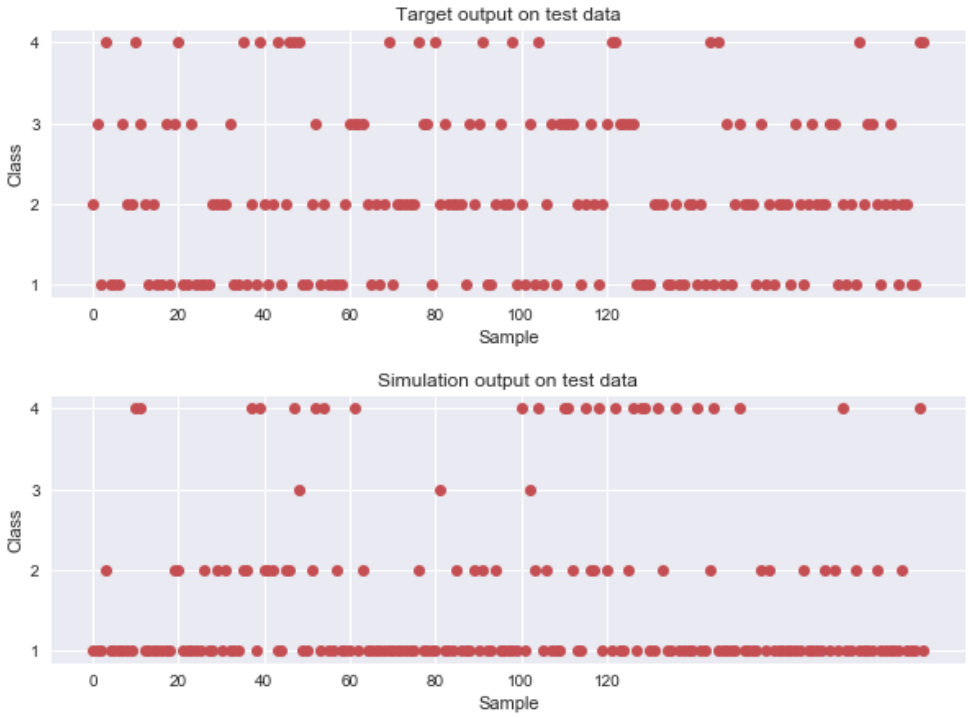
F1-skóre v reporte knižnice *Sklearn* v prípade *multiclass* klasifikácie predstavuje vážený priemer F1-skóre každej triedy. Je počítané na základe hodnôt *precision* a *recall*. Mierne vyššia *precision* naznačuje mierne nižšie množstvo nesprávne pozitívne klasifikovaných prvkov. Nižší *recall* naznačuje viac nesprávne negatívne označených prvkov, v tomto prípade predovšetkým u prvého čísla z dvojčísla určujúceho výslednú triedu, z čoho vyplýva, že sa to dotkne najmä 3. a 4. triedy.

	precision	recall	f1-score	support
0	0.54	0.24	0.33	62
1	0.57	0.40	0.47	89
micro avg	0.56	0.34	0.42	151
macro avg	0.55	0.32	0.40	151
weighted avg	0.56	0.34	0.42	151
samples avg	0.19	0.21	0.19	151

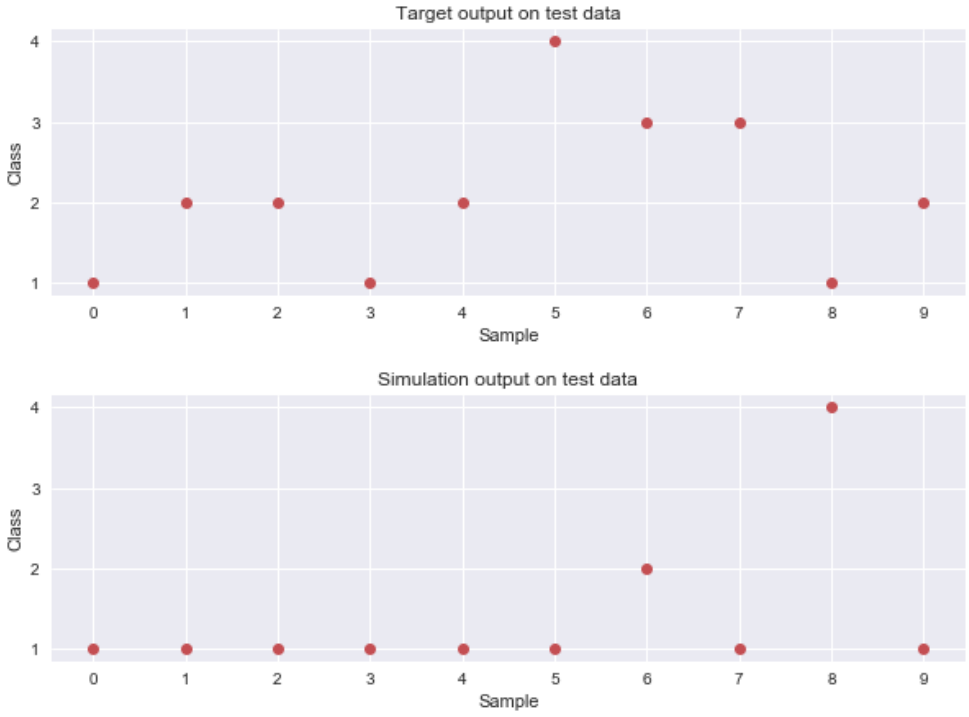
Success: 39.49 %

Obrázok 17: Report predikcie testovaných dát neurónovej siete

Z hľadiska trendu medzi klasifikovanými dátami, je vidno, že najhoršie na tom vyzerá byť 3. trieda, ktorá bola neurónovou sieťou odhalená málokrát. Pri náhodných 10 vzorkoch je vidno, že existuje určitý pokus o podobný sklon krivky, ale málo dvíhajúci sa na y-osi, čo korešponduje s výsledkami F1-skóre.



Obrázok 18: Porovnanie klasifikovaných a ich správnych hodnôt tried



Obrázok 19: Porovnanie desiatich náhodných klasifikovaných a ich správnych hodnôt tried

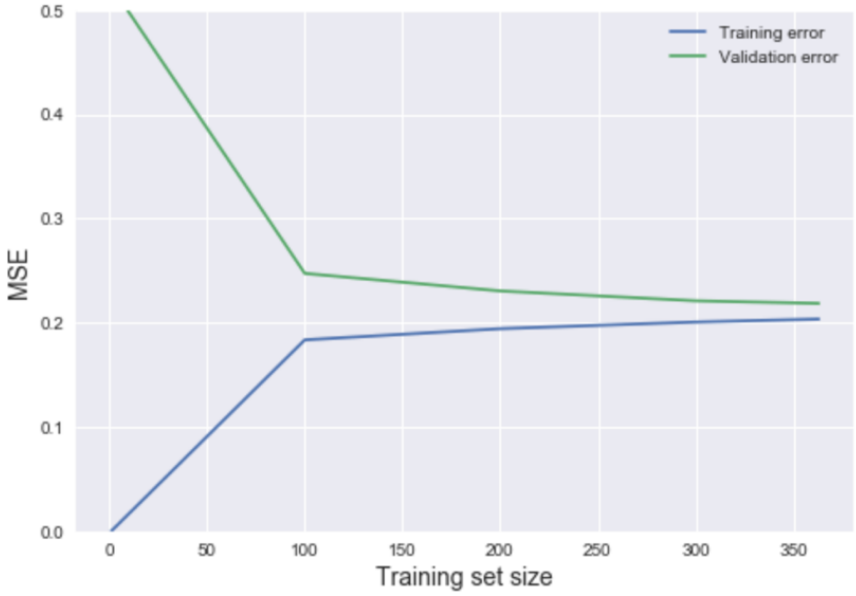
3. neurónová sieť – estimátor lineárnej regresie najmenších štvorcov

V poradí tretia neurónová sieť – estimátor lineárnej regresie dosahovala najlepšej úspešnosti pri pomere rozdelenia dát 70/30 a teda 70% trénovacích dát a 30% testovacích v hodnote 35,38%.

Pomery	70/30	80/20	90/10
Úspešnosť testovania	35,38 %	32,31 %	32,31 %

Tabuľka 8: Pomery trénovacích, validačných a testovacích dát

Krivka trénovacích aj validačných dát má na základe MSE relatívne nízku chybovosť v porovnaní s predchádzajúcou nerónovou sieťou, takže aj *bias* je lepší. Taktiež *variancia* je nízka, takže krivka sa približuje k ideálnemu tvaru pre žiadny sklon k preučeniu a podučeniu. Taktiež môžeme vidieť, že od 100 prvkov sa jej chybovosť značne zmení. Tento prechod by v ideálnom prípade mal byť menej nárazový a trénovacia krivka by sa mala iterovať postupne.



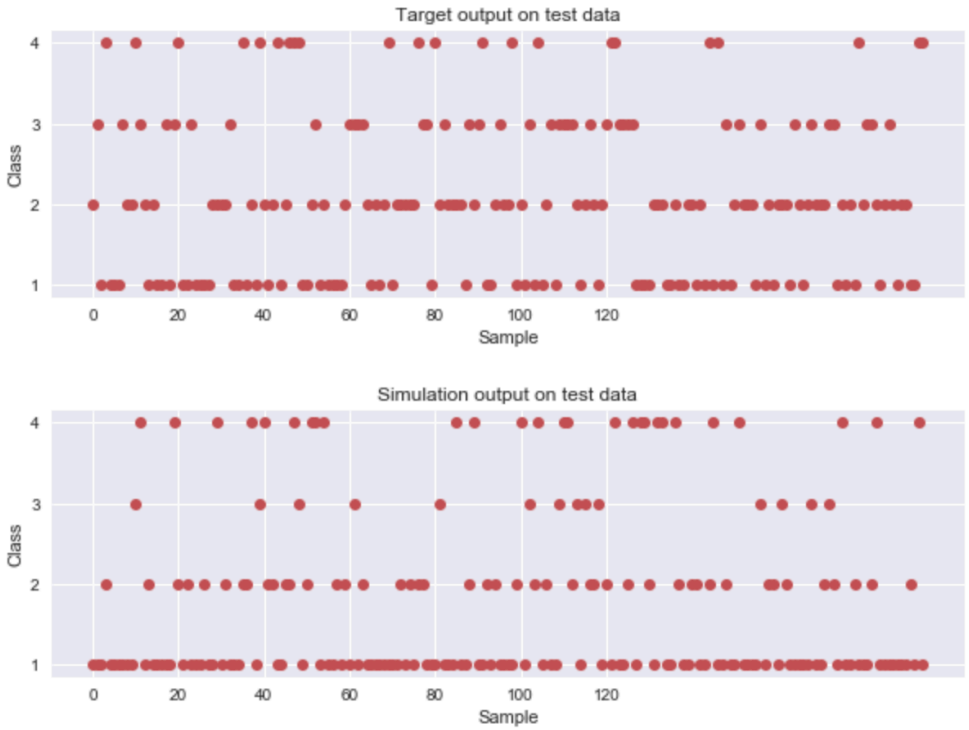
Obrázok 20: Učiaca krivka pre pomer dát 70/30 [%]

Hodnota F1-skóre je nižšia ako v prípade prvej neurónovej siete. Hodnota *precision* je opäť mierne vyššia o málo nižšia ako v predchádzajúcom prípade. Avšak, stúpol *recall*, ktorý naznačuje nesprávne negatívne označených prvkov.

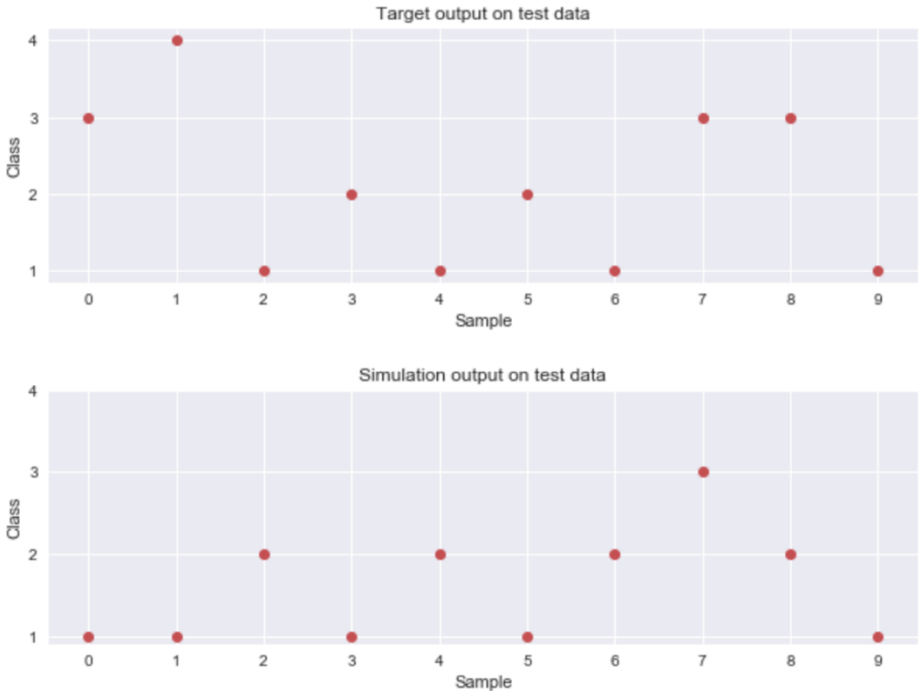
	precision	recall	f1-score	support
0	0.49	0.32	0.39	62
1	0.50	0.40	0.45	89
micro avg	0.50	0.37	0.42	151
macro avg	0.49	0.36	0.42	151
weighted avg	0.49	0.37	0.42	151
samples avg	0.22	0.24	0.22	151

Obrázok 21: Report predikcie testovaných dát neurónovej siete

Na základe nasledujúceho grafu možno sledovať, že trend medzi klasifikovanými dátami je lepší v porovnaní s predchádzajúcou neurónovou sieťou - tretia trieda už bola klasifikovaná častejšie. Každopádne v porovnaní s predchádzajúcimi dvoma výsledkami dokázala v tomto prípade neurónová sieť klasifikovať najrozmanitejšie výsledky. Porovnanie pri desiatich náhodných prvkov vyzerá obdobne úspešne ako v predchádzajúcej neurónovej sieti.



Obrázok 22: Porovnanie klasifikovaných a ich správnych hodnôt tried

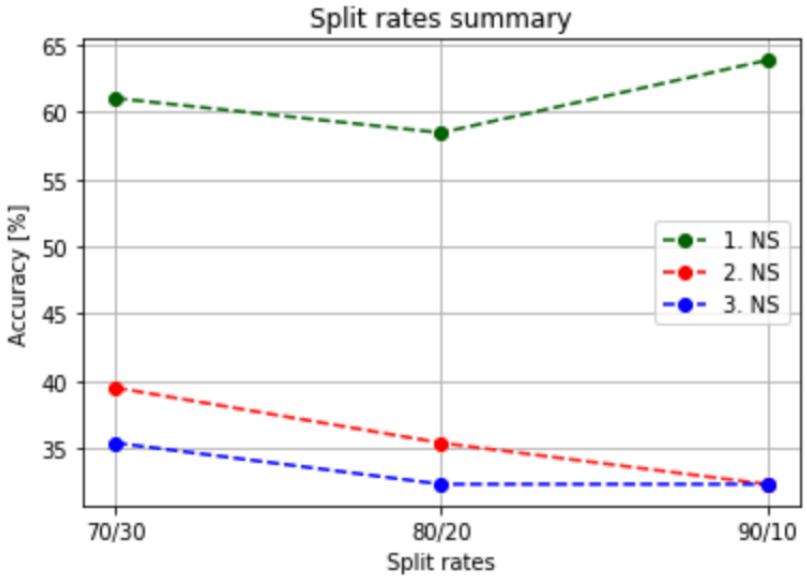


Obrázok 23: Porovnanie desiatich náhodných klasifikovaných a ich správnych hodnôt tried

3.3. Vyhodnotenie celkovej úspešnosti

Napriek tomu, že je na prvý pohľad prívetivejším grafom graf tretej neurónovej siete, mala druhá neurónová sieť mierne vyššiu úspešnosť. Zo všetkých troch neurónových sietí mala najlepšiu úspešnosť sieť knižnice *Neurolabu* s doimplementovanou validáciou. Je možné, že táto sieť prešla v súčte s epochami a iteráciami vyšším tréningom a to mohlo ovplyvniť jej výsledky. Taktiež vnútorná implementácia tejto siete bude určite mierne odlišná.

Z hľadiska úspešností vyhodnocovania dát mali všetky neurónové siete pomerne slabé výsledky a neprekročili hranicu 65 % úspešnosti. Toto môže byť spôsobené vysokým množstvom vlastností (angl. *features*) a nízkym počtom dát. Taktiež je možné, že dáta majú v rámci jednej cieľovej triedy (angl. *targetu*) vysokú rôznorodosť vlastností.



Obrázok 24: Prehľad úspešnosti jednotlivých neurónových sietí podľa pomerov rozdelenia dát

4. Záver

Cieľom práce bolo vytvorenie neurónovej siete pre riešenie klasifikačného problému súvisiaceho s konzumáciou alkoholu medzi študentami stredných škôl. Neurónová sieť bola namodelovaná vo viacerých podobách prostredníctvom dvoch knižníc jazyka *Python*. Pred implementáciou boli dáta upravené a normalizované, aby mohla neurónová sieť s údajmi pracovať. Následne boli znázornené závislosti medzi atribútmi a triedami.

V rámci implementácie sme pracovali s viacerými hodnotami *lambda* a pomermi rozdelenia dát, aby sme dosiahli čo najlepší výsledok úspešnosti neurónovej siete. Tento výsledok bol najlepší pre neurónovú sieť č.1, avšak všetky dosahovali slabých úspešností pravdepodobne spôsobené malým množstvom dát. Skúsenosť pri tvorbe tejto práce bola prívetivá a vo výsledku nám umožnila lepšie pochopenie princípov vytvárania neurónových sietí.

5. Zdroje

STACK EXCHANGE INC. *How to choose the number of hidden layers and nodes in a feedforward neural network?* [online, cit. 2020-06-06]. Dostupné z: https://stats.stackexchange.com/questions/181/how-to-choose-the-number-of-hidden-layers-and-nodes-in-a-feedforward-neural-netw

SCIKIT-LEARN DEVELOPERS (BSD LICENSE). *Sklearn.neural_network.MLPClassifier* [online, cit. 2020-06-06]. Dostupné z: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

6. Prílohy

A. Popisy použitých atribútov

sex – pohlavie študenta
0: žena, 1: muž

age – vek študenta
v rozmedzí 15–22

family_size – počet rodinných príslušníkov
O: väčšie alebo rovné 3, 1:– menej alebo rovné 3

parents_status – vzťah rodičov
0: nebývajú spolu, 1: bývajú spolu

mothers_education – vzdelanie matky
0: žiadne, 1: základné (do 4 triedy), 2: základné (5. – 9. trieda), 3: stredoškolské, 4: vysokoškolské

father_education – vzdelanie otca
0: žiadne, 1: základné (do 4 triedy), 2: základné (5. – 9. trieda), 3: stredoškolské, 4: vysokoškolské

mothers_job – zamestnanie matky
0: žiadne, 1: zdravotníctvo, 2: ostatné, 3: služby, 4: učiteľ

fathers_job – zamestnanie otca
0: žiadne, 1: zdravotníctvo, 2: ostatné, 3: služby, 4: učiteľ

guardian – zákonný zástupca
0: matka, 1: otec, 2: ostatné, 3: ostatné

study_time – čas strávený študovaním týždenne
1: <1;2) hodiny, 2: <2-5) hodín, 3: <5-10) hodín, 4: <10; viac) hodín

failures – počet nezvládnutých predmetov
v rozmedzí 1–4

school_support – podpora v podobe školských doučovaní
1: áno, 0: nie

family_support – podpora v podobe doučovaní doma
1: áno, 0: nie

activities – extra aktivity
1: áno, 0: nie

higher_education – snaha dosiahnuť vyššie vzdelanie
1: áno, 0: nie

internet – prístup k internetu
1: áno, 0: nie

relationship – romantický vzťah
1: áno, 0: nie

family_relationship– kvalita vzťahu s rodičmi
v rozmedzí od 1(veľmi zlý) po 5 (veľmi dobrý)

free_time – množstvo voľného času po škole
v rozmedzí od 1 (veľmi málo) po 5 (veľmi veľa)

go_out– ako často chodí študent von
v rozmedzí od 1 (veľmi málo) po 5 (veľmi veľa)

health – zdravotný status
v rozmedzí od 1 (veľmi zlý) po 5 (veľmi dobrý)

absences – počet absencí v škole
v rozmedzí 0–93

final_grade – výsledná známka
v rozmedzí 0–20

alcohol_consumtion – miera konzumácie alkoholu
v rozmedzí od 1 (veľmi málo) po 5 (veľmi veľa)

B. Súbory

- Zdrojový kód pre merania rôznych hodnôt alpha
- Zdrojový kód pre merania analýzy chybových vzorkov
- Zdrojový kód pre merania rôznych pomerov trénovacích, validačných a testovacích dát
- Zdrojový kód pre súhrnné tabuľky