SUNS – Zadanie 1

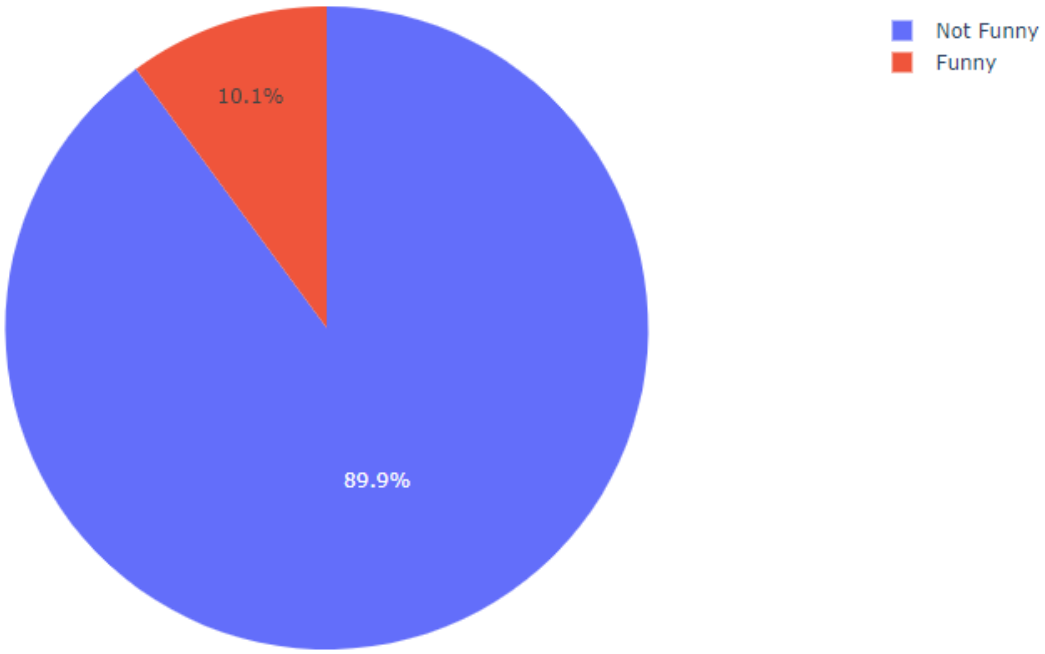
Marek Dráb

97757

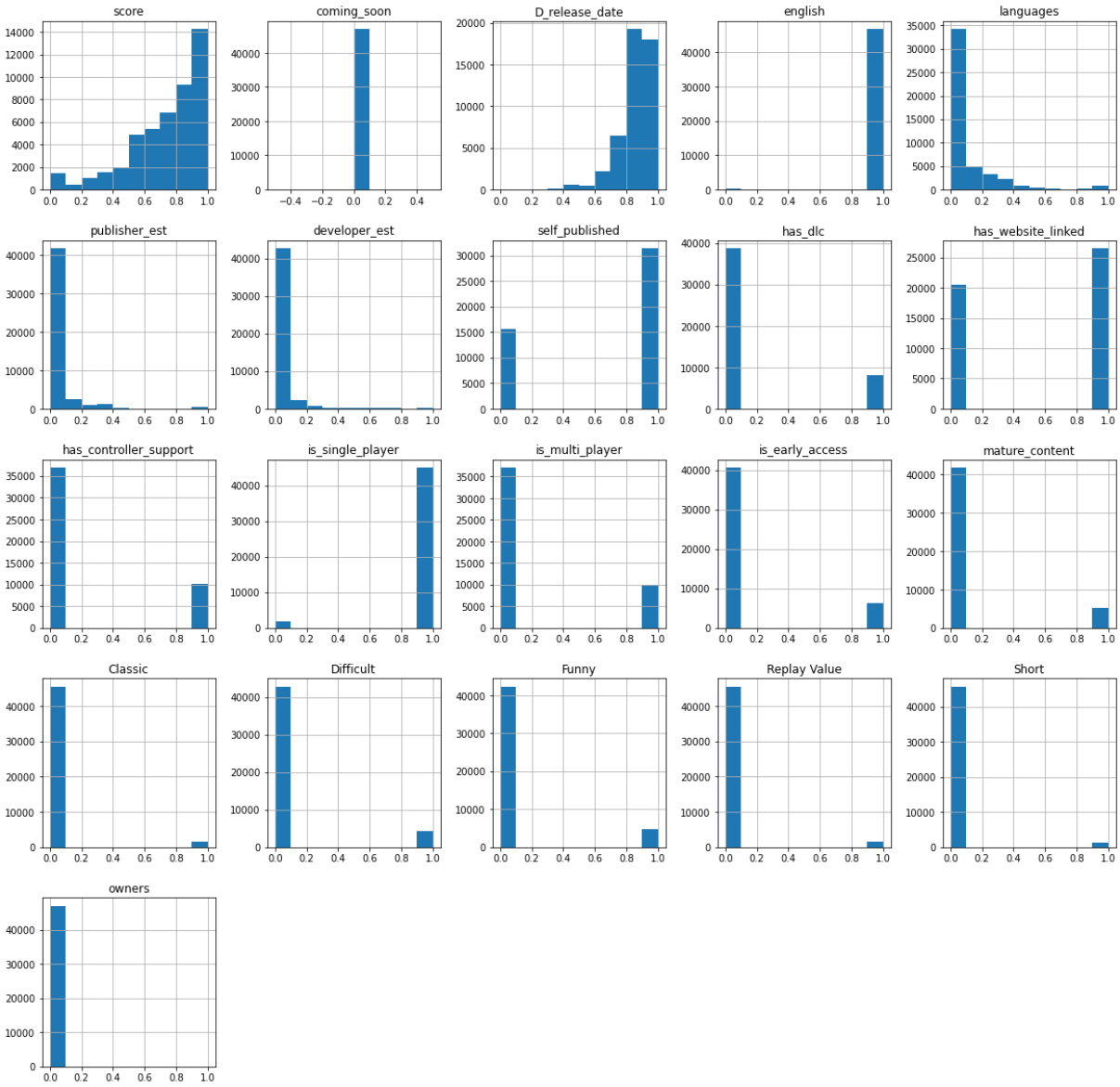
Dáta

Dáta sú načítavané z csv súborov. Obsahujú Nan a null hodnoty, ktoré odstraňujem pomocou funkcie dropna z pandasu. Počet sa z pôvodných vyše 62000 záznamov zredukoval na 47055 riadkov.

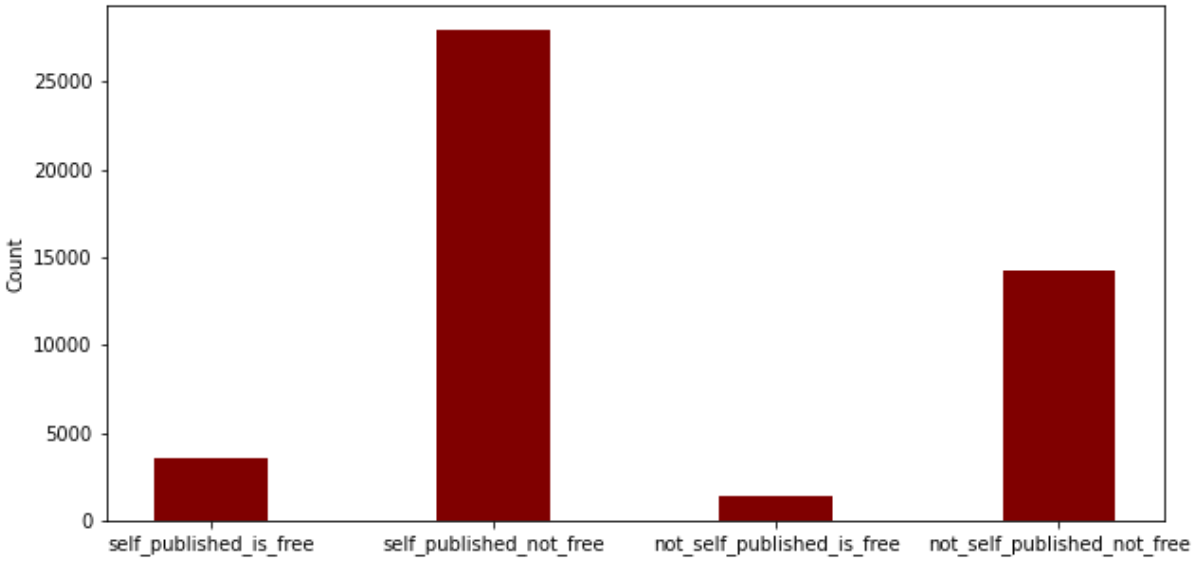
EDA



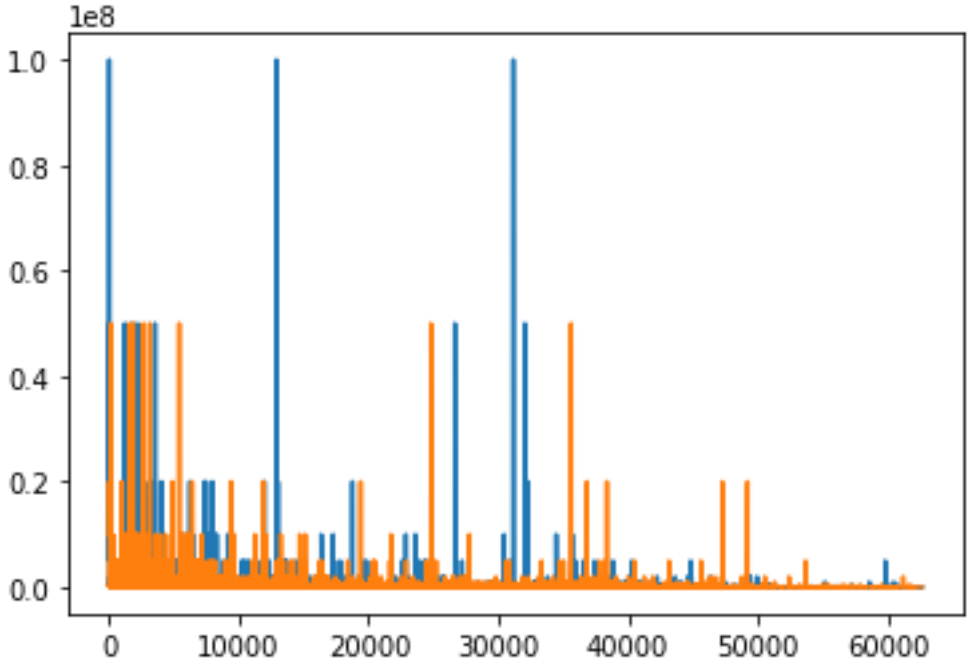
Obrázok 1 Pomer Funny a Not Funny



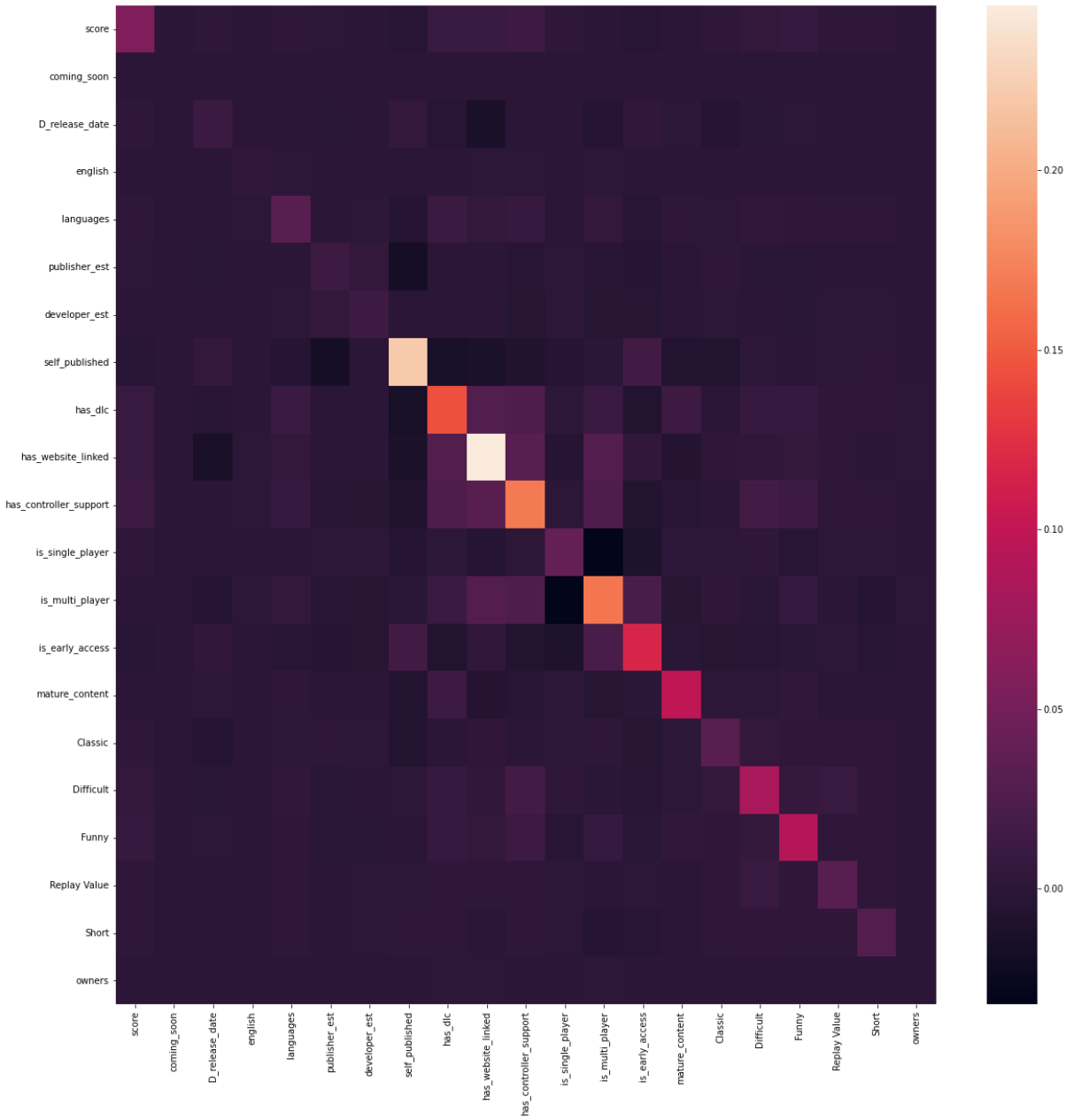
Obrázok 2 Histogram škálovaných dát



Obrázok 3 Súvis medzi self_published a is_free



Obrázok 4 Počet majiteľov self_published a not_self_published hier



Obrázok 5 Konvolučná matica po škálovaní

Získanie roku z dátumu

df['D_release_date'] = pd.DatetimeIndex(df['D_release_date']).year

Pomocou funkcie DatetimeIndex z knižnice Pandas sa dá vytiahnuť jednotlivé časti dátumu, teda aj rok.

Počet majiteľov hry

df['D_owners'] = pd.Series(df['D_owners'], dtype="string")

tmp = df['D_owners'].str.split(" .. ", expand=True)[1].str.replace(',', '')

tmp = tmp.astype(int)

df['owners'] = tmp.copy()

Stĺpec *D_owners* je typu Series. Treba pretypovať na string a je s ním možné narábať ako so stringom. Následne ho splitnem, vezmem väčšiu hodnotu a z tej odstránim znaky ,,‘. Teraz je možné hodnotu pretypovať na int.

Stĺpce

Vyhodil som všetky D_ stĺpce okrem D_release_date – ten má prepísanú hodnotu na rok. Z ostatných stĺpcov som odstránil:

- Positive a negative – vysoká korelácia hodnôt medzi sebou
- Ccu – nedáva mi zmysel to tam nechať, veľa hráčov môžu mať platené aj neplatené hry
- Addictive
- Beautiful
- Masterpiece
- Well-Written
- Lore-Rich
- Epic
- Emotional
- Cult Classic
- Competetive

Všetky z týchto stĺpcov mali menej ako 1000 hodnôt True.

Škálovanie

Na škálovanie bol použitý MinMaxScaler, ktorý hodnoty upravil do rozmedzia 0 až 1.

Trénovacie dáta

Hodnoty pred škálovaním:

score	2.400225e-01
D_release_date	2.576946e+00
languages	5.004820e+00
publisher_est	4.584564e+01
developer_est	1.476543e+01
owners	1.459761e+06

Tabuľka 1 Priemerné hodnoty v trénovacej množine

score	2.213345e-01
D_release_date	2.466754e+00
languages	4.536851e+00
publisher_est	3.811713e+01
developer_est	9.474395e+00
owners	7.116430e+06

Tabuľka 2 Priemerné hodnoty v testovacej množine

Hodnoty po škálovaní:

score	0.240023
D_release_date	0.103078
languages	0.178744
publisher_est	0.113761
developer_est	0.107777
owners	0.014601

Tabuľka 3 Priemerné hodnoty v trénovacej množine

score	0.221334
D_release_date	0.117464
languages	0.162030
publisher_est	0.094583
developer_est	0.086921
owners	0.035586

Tabuľka 4 Priemerné hodnoty v testovacej množine

Trénovanie

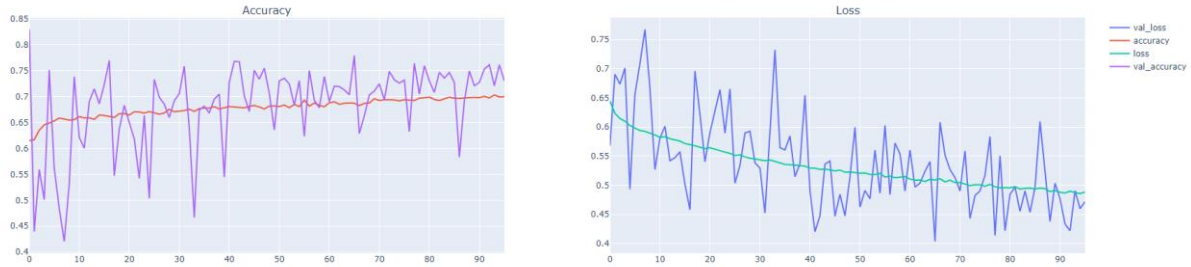
Architektúra modelu

- Viacvrstvová neurónová sieť
- Skryté vrstvy relu
- Rozdelenie 64 – 32 – 64
- Výstupná vrstva sigmoid
- 1 výstupný neurón
- Learning rate = 0.002
- Adam optimizer
- 5% validačná množina
- 200 epoch
- Veľkosť batchu = 64
- Vyrovnanie rozdielneho pomeru hodnôt pomocou class_weight

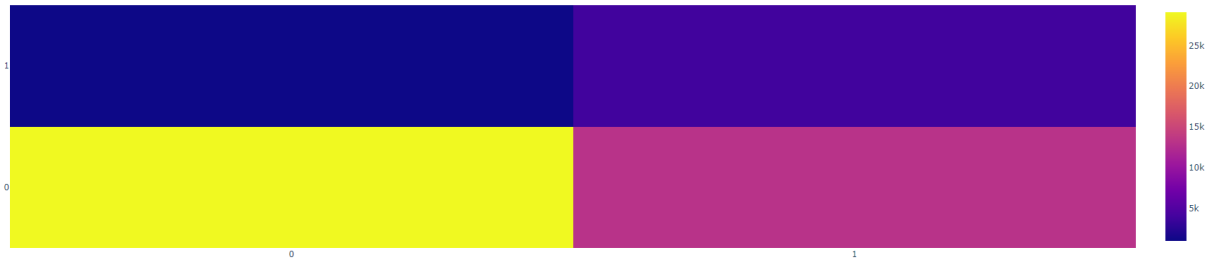
```
class_weights = class_weight.compute_class_weight(
                                        class_weight = "balanced",
                                        classes = np.unique(y_train),
                                        y = y_train
                                    )
class_weights = dict(zip(np.unique(y_train), class_weights))
class_weights
```

```
{False: 0.5582247846821838, True: 4.793704156479218}
```

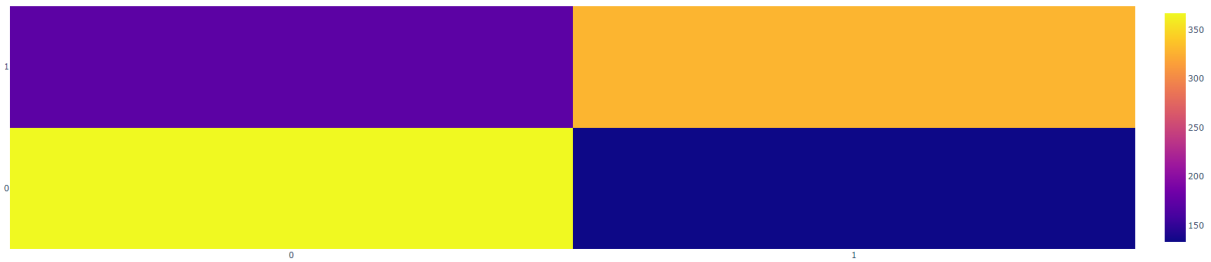
Obrázok 6 Class weights



Obrázok 7 Priebeh trénovania



Obrázok 8 Konfúzna matica trénovacej množiny

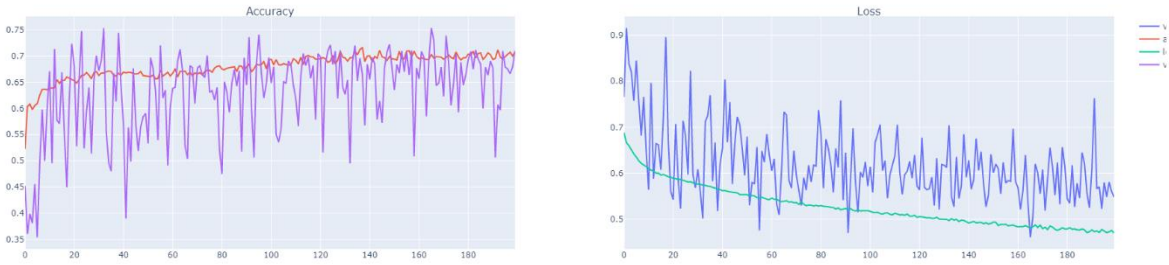


Obrázok 9 Konfúzna matica testovacej množiny

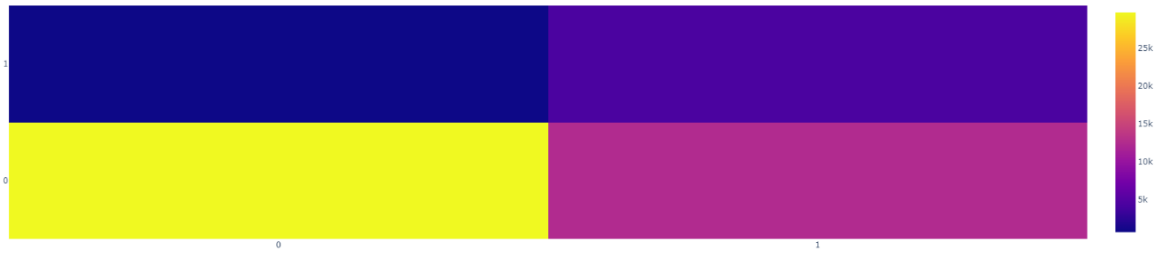
Pretrénovanie

- Viac vrstiev ako pri predošlom modeli
- Relu – 256
- Relu – 128

- TanH – 512
- TanH – 128
- Softmax – 64
- 200 epoch bez early stoppingu



Obrázok 10 Priebeh trénovania - pretrénovanie

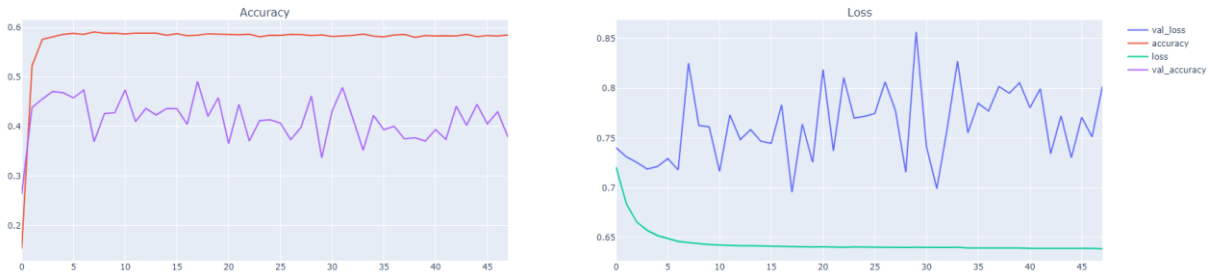


Obrázok 11 Konfúzna matica – trénovacia množina

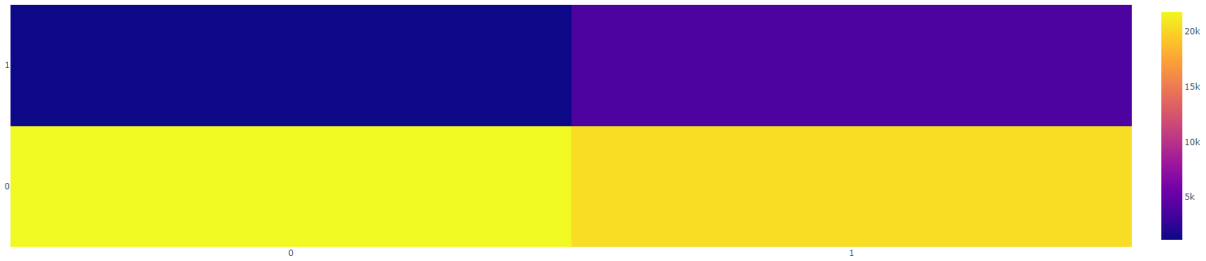
Obrázok 12 Konfúzna matica - testovacia množina

Podtrénovanie

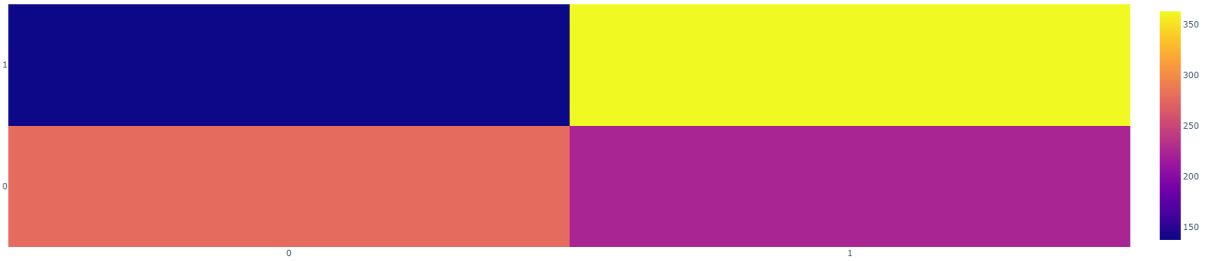
- Relu – 4
- Softmax – 4
- Learning rate = 0,001



Obrázok 13 Priebeh trénovania – podtrénovanie



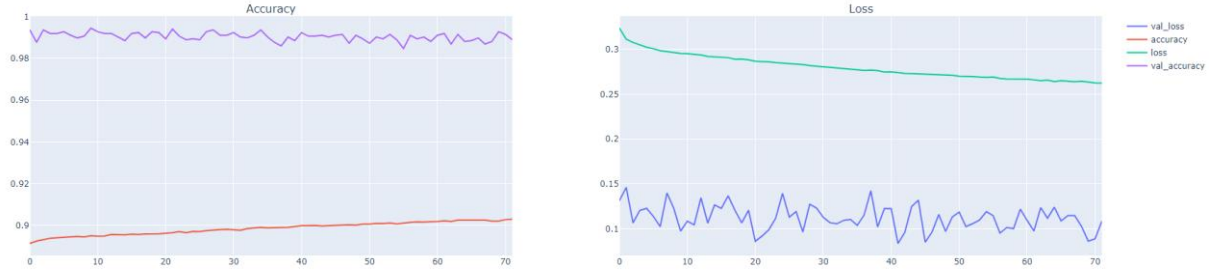
Obrázok 14 Konfúzna matica - trénovacia množina



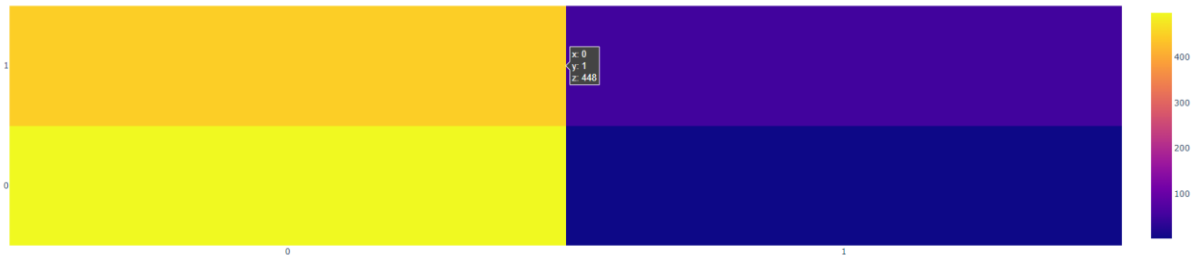
Obrázok 15 Konfúzna matica – testovacia množina

Bez použitia váh

Model je rovnaký ako v najlepšom nastavení.



Obrázok 16 Priebeh trénovania - bez použitia váh



Obrázok 17 Konfúzna matica - bez použitia váh