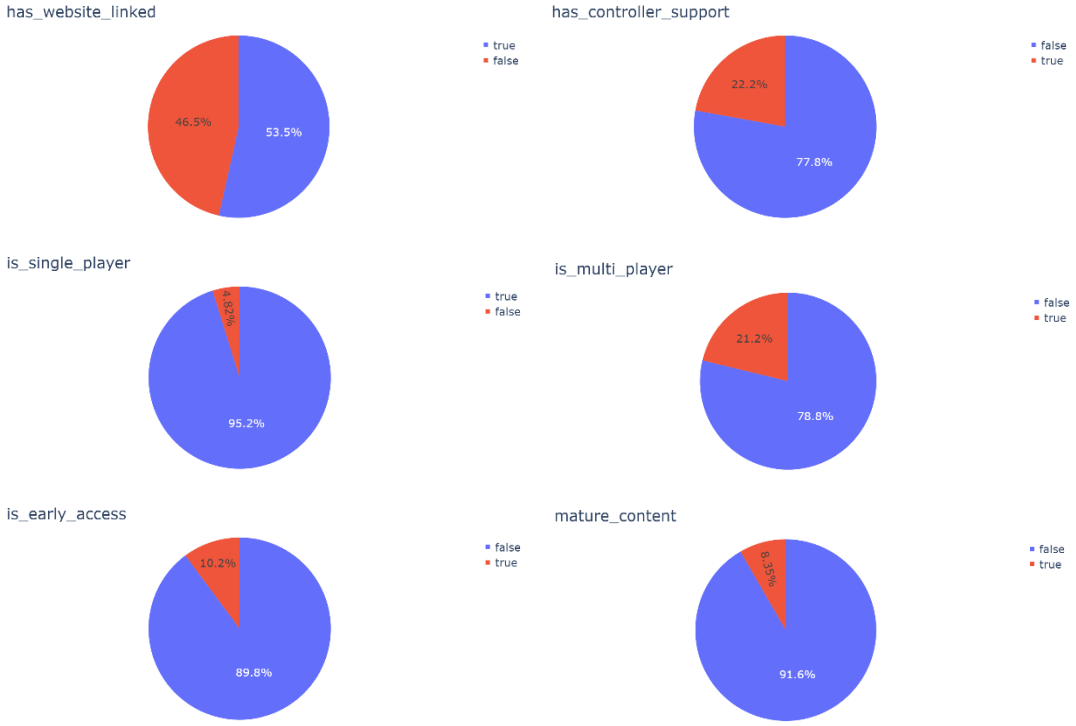
# I-SUNS: Zadanie č.1

Neurónové siete

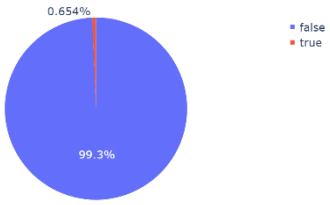
Tomáš Minárik

## Príprava dát

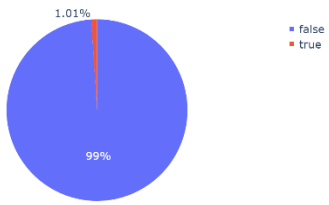
- Najprv sme si načítali dáta pomocou funkcie read_csv. Po načítaní dát sme ako prvé skontrolovali riadky kde sa nachádzajú nulové hodnoty. Zistili sme že aspoň jedna nulová hodnota sa nachádza v 15 841 z 62 869 riadkov.
- Rozhodli sme sa tieto riadky odstrániť keďže neskôr použijeme undersampling.
- Ďalej sme sa rozhodli odstrániť stĺpce s id a menami - D_appid, D_name, D_developer, D_publisher.
- Z hodnotení sme sa rozhodli ponechať len stĺpec score a odstrániť stĺpce positive, negative, D_reviews
- Zo stĺpca D_owners sme dostali počet majiteľov tak že sme spravili priemer z hornej a dolnej hranice.
- Zo stĺpca D_release_date sme vybrali rok.
- Stĺpec D_genre sme odstránili keďže žánre budeme ďalej vyberať z tagov.
- Ďalej sme sa pozreli na stĺpce ktoré obsahovali hodnoty True a False .
- Odstránili sme stĺpec coming_soon lebo obsahoval len hodnoty False.



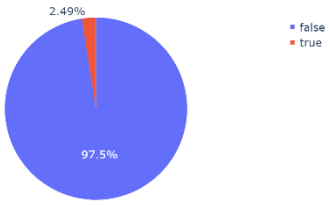
Addictive



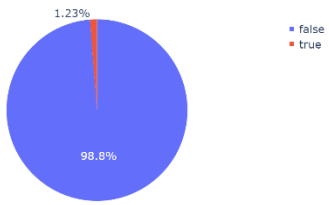

Beautiful



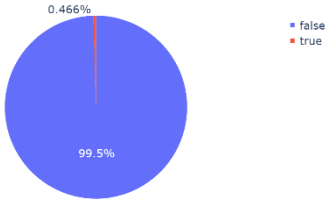

Classic



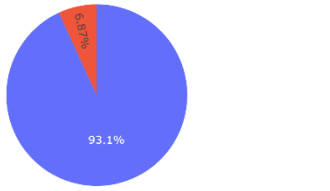

Competitive



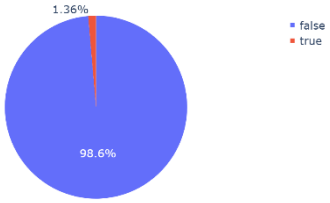

Cult Classic



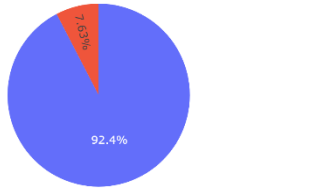

Difficult



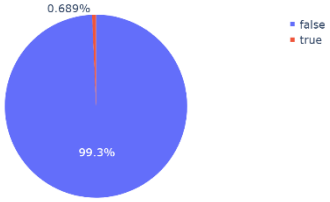

Emotional



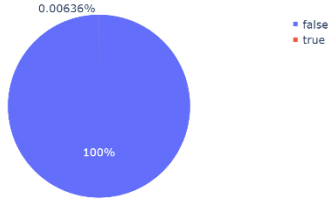

Funny



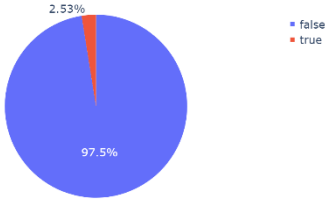

Lore-Rich



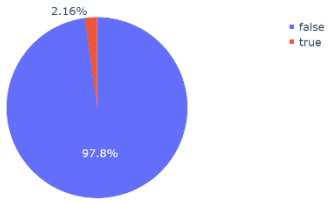

Masterpiece

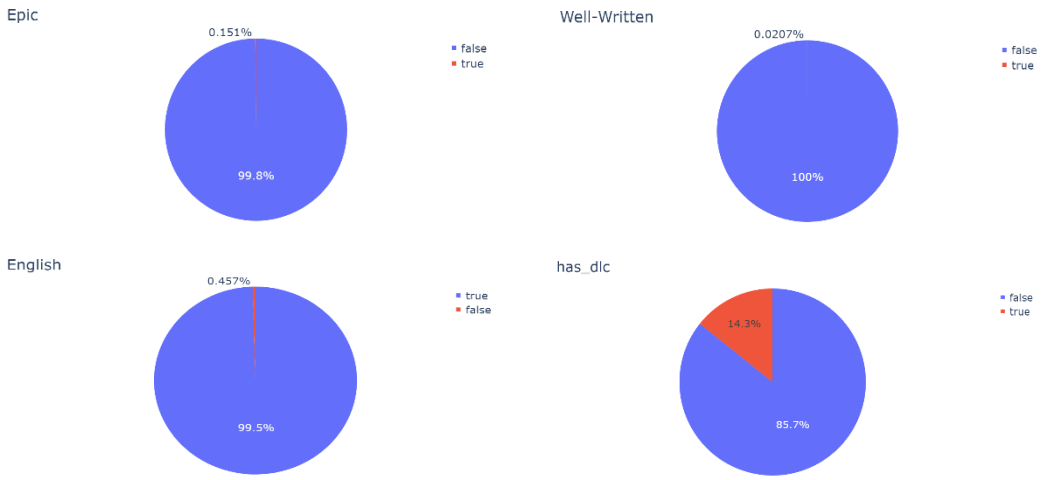



Replay Value




Short






- Z grafov môžeme vidieť že väčšina stĺpcov obsahuje viac ako 90% len hodnotu False a preto sme sa rozhodli ponechať len stĺpce has_website_linked, is_multi_player, has_controller_support, has_dlc.
- Ďalej sme zo stĺpca D_tags získali tagy, ktoré sa vyskytujú najčastejšie.

| Tags | |
|---|---|
| Indie | 32319 |
| Singleplayer | 21884 |
| Action | 21703 |
| Casual | 20747 |
| Adventure | 20158 |
| 2D | 11606 |
| Strategy | 10290 |
| Simulation | 10003 |
| RPG | 8693 |
| Puzzle | 8134 |
| Atmospheric | 7586 |
| Early Access | 6366 |
| Story Rich | 6038 |
| Pixel Graphics | 5990 |
| Multiplayer | 5774 |
| Colorful | 5310 |
| Arcade | 5288 |
| 3D | 5263 |
| Cute | 5103 |
| First-Person | 4940 |

- Z vyššie uvedeného zoznamu sme vybrali 8 najčastejšie sa vyskytujúcich žánrov a pre každý sme vytvorili vlastný stĺpec(Indie, Singleplayer, Action, Casual, Adventure, 2D, Strategy, Simulation, RPG).
- Zo stĺpcov, ktoré informovali o vydavateľovi sme sa rozhodli ponechať stĺpce publisher_est, self_published a odstrániť stĺpec developer_est.

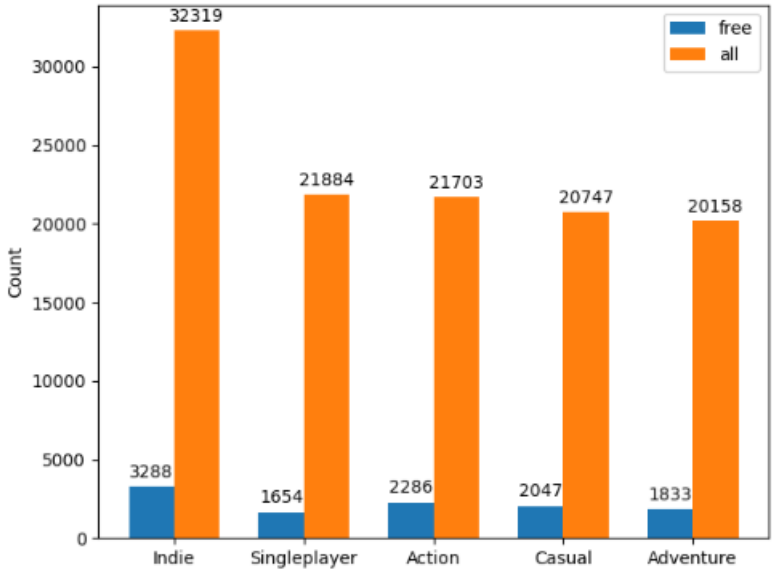
- Stĺpec ccu sme sa rozhodli odstrániť keďže máme stĺpec s počtom majiteľov a zároveň 99.97% nemá viac ako 100 hráčov.
- Na záver sme normalizovali stĺpce s číselnými hodnotami(D_owners, D_release_date, languages, publisher_est). V nasledujúcom obrázku som zobrazené priemery pred a po použitím MinMaxScalera.

```
Before
Owners:  122628.41355860163
Date:  2018.2569121241102
languages:  3.6380405907980022
publishers:  17.478588885346934
After
Owners:  0.0015019124357727915
Date:  0.8502764849644099
languages:  0.09421573538564294
publishers:  0.04088979872294524
```

- Rovnaké úpravy boli spravené aj na testovacích dátach

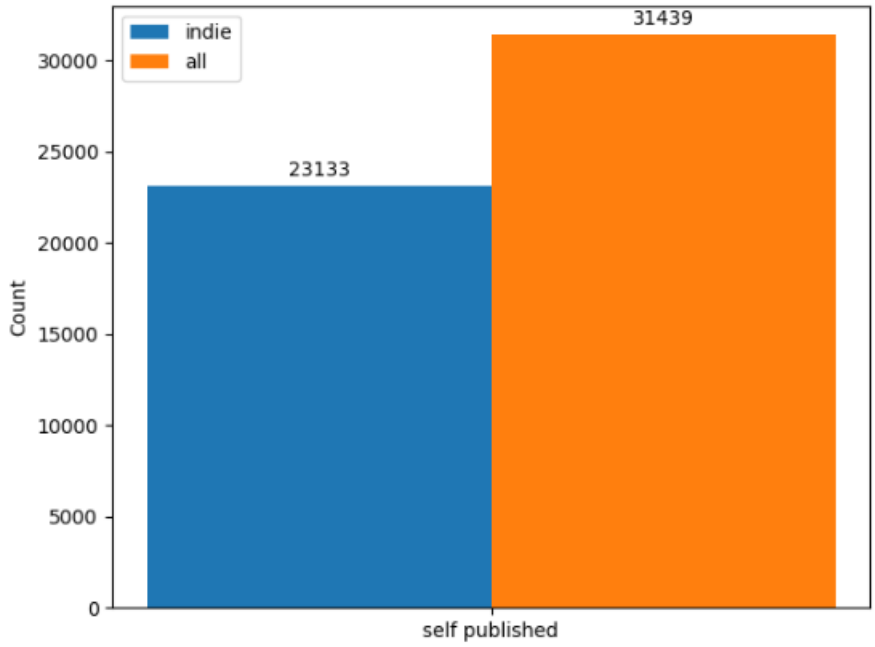
## EDA

**Top žánre a is_free**



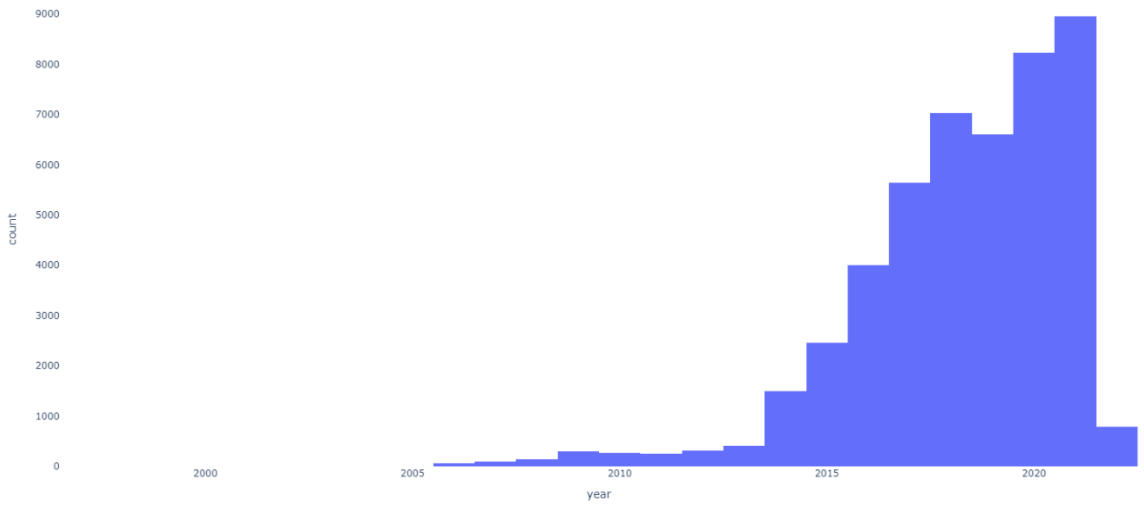

Z grafu môžme vidieť že najlepší pomer hier zadarmo k plateným hrám má žáner action pri ktorom je to približne 10.5%. Najhoršie je na tom žáner singleplayer z ktorého je približne 7.6% hier zadarmo.

**Self published a Indie**



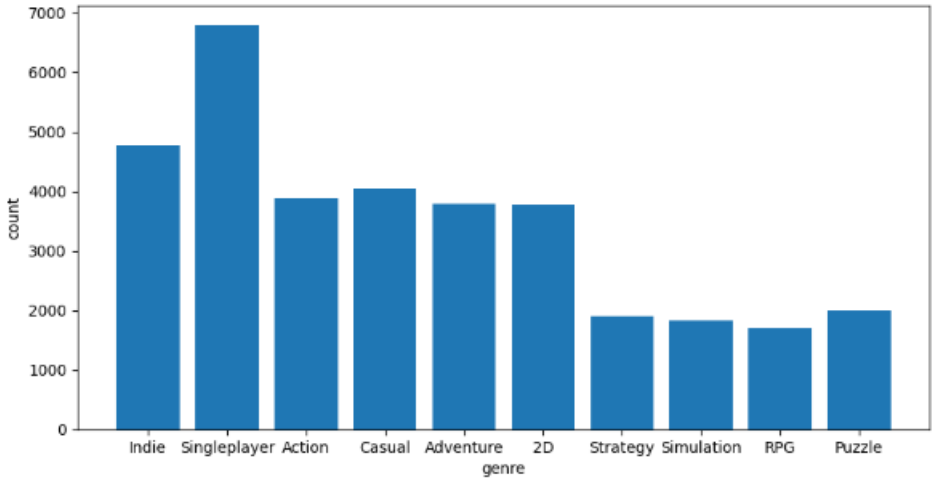

Z grafu môžme vidieť že väčšina hier, ktoré sú self published patria pod žáner indie. Čo sa dalo predpokladať keďže indie znamená nezávislá a ide o hry tvorené jednotlivcom alebo malým tímom.

**Počet vydaných hier podľa rokov**




Z grafu môžeme vidieť že takmer každý rok je vydaných viac hier ako v predchádzajúcom roku. Zatiaľ najviac vydaných hier bolo v roku 2021 a to 8954.

**Top herné žánre v roku 2021**

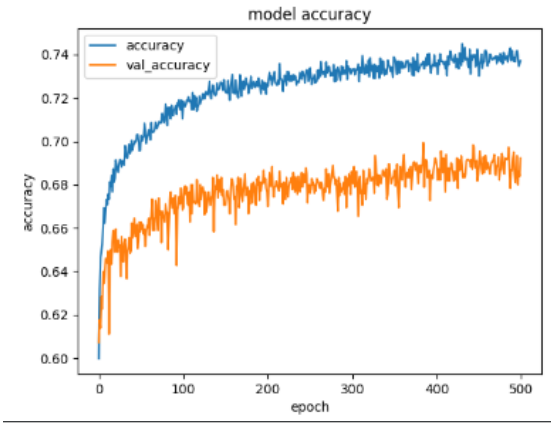


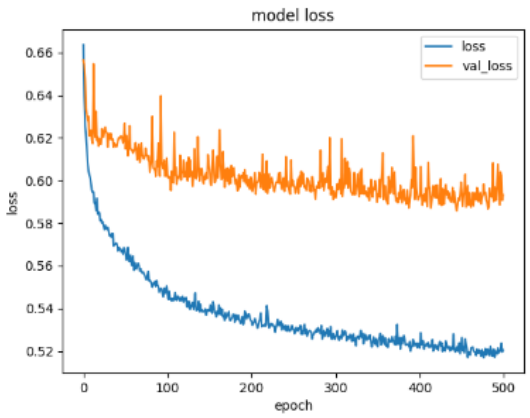

V roku 2021 bolo najviac hier vzdaných ako singleplayer na druhom mieste boli indie a na treťom casual.

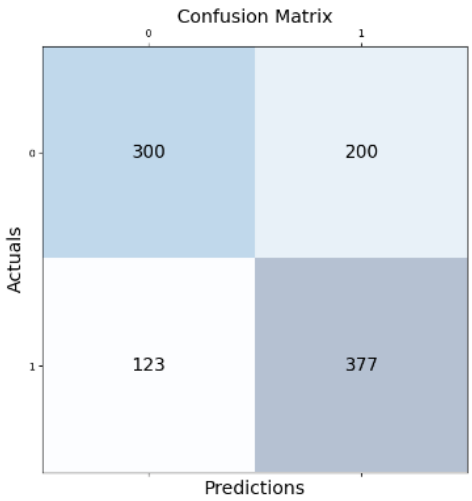
## Trénovanie

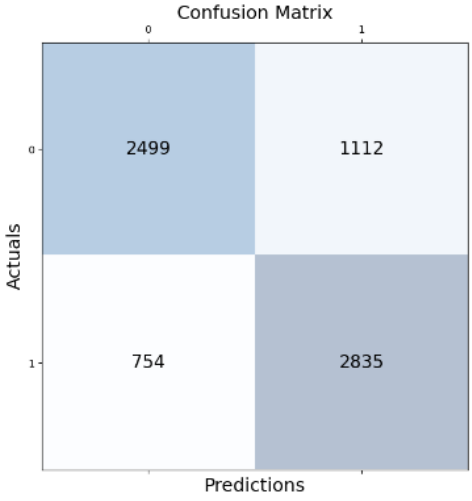
### Sieť 1

- Vstupné parametre(20) - score,D_owners,D_release_date,languages,publisher_est,self_published,has_dlc,has_website_linked,has_controller_support,is_multi_player,is_early_access,Indie,2D,Action,Casual,Adventure,Singleplayer,Strategy,Simulation,RPG
- Rozdelenie:
    - 4500 zadarmo
    - 4500 platené
- Validačné dáta – 20% z trénovacích
- 1 skrytá vrstva s 20 neurónmi a aktivačnou funkciou relu
- 1 výstupný neurón, aktivačná funkcia sigmoid
- Kriteriálna funkcia Binary Cross Entropy
- Solver – Adam s learning rate 0.01
- Batch size = 100
- Počet epoch = 500



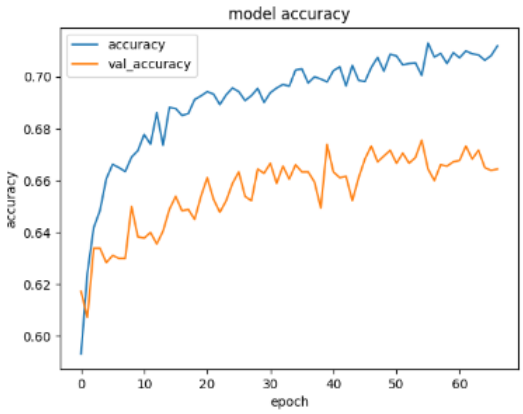



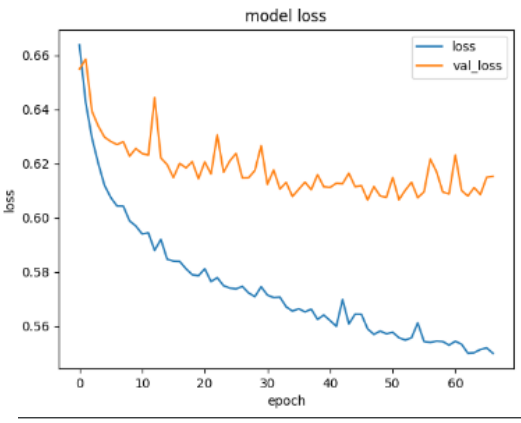

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.71 | 0.60 | 0.65 | 500 |
| 1.0 | 0.65 | 0.75 | 0.70 | 500 |
| accuracy | | | 0.68 | 1000 |
| macro avg | 0.68 | 0.68 | 0.68 | 1000 |
| weighted avg | 0.68 | 0.68 | 0.68 | 1000 |

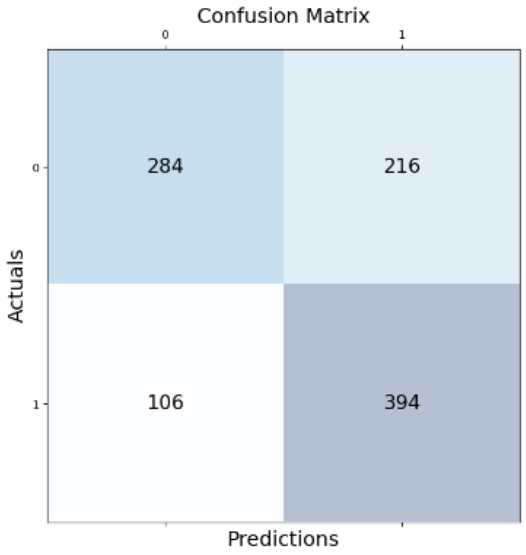
- Úspešnosť na trénovacích dátach je 73% a validačných je 69%
- Úspešnosť na testovacích  dátach je 68%.

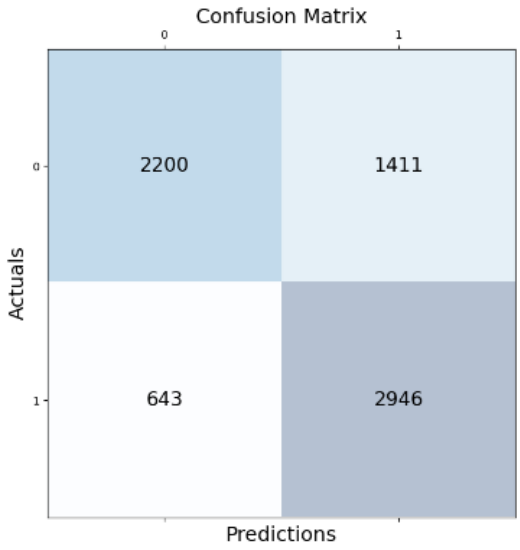
## Sieť 2

- Rovnaká sieť ako sieť 1
- Pridaný EarlyStopping s patience 20



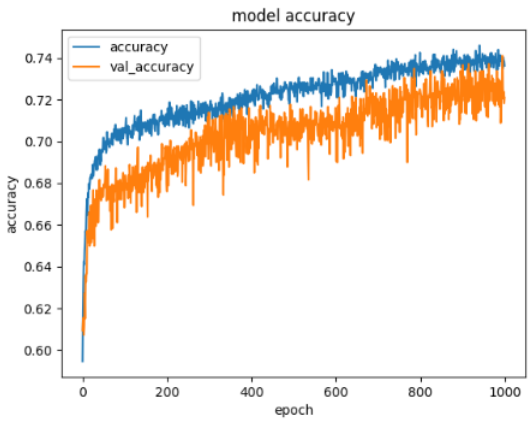

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.73 | 0.57 | 0.64 | 500 |
| 1.0 | 0.65 | 0.79 | 0.71 | 500 |
| accuracy | | | 0.68 | 1000 |
| macro avg | 0.69 | 0.68 | 0.67 | 1000 |
| weighted avg | 0.69 | 0.68 | 0.67 | 1000 |

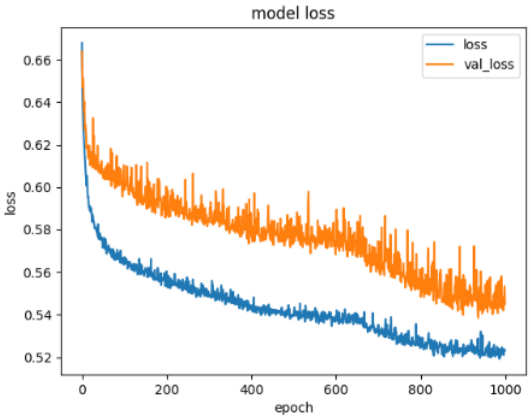
- Úspešnosť na trénovacích dátach je 71% a na validačných 66%.
- Úspešnosť na testovacích dátach je 69%.
- Trénovanie bolo ukončené po 67. epoche.

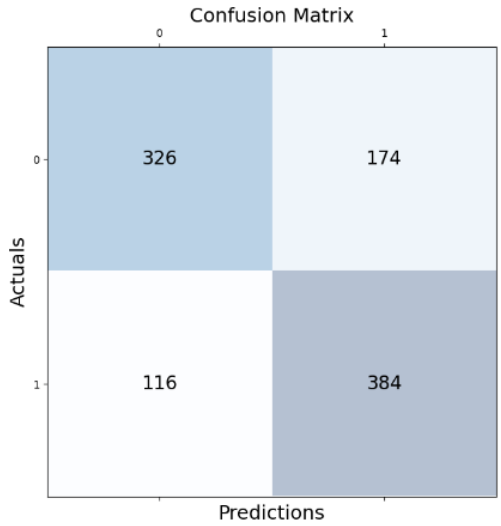
## Sieť 3

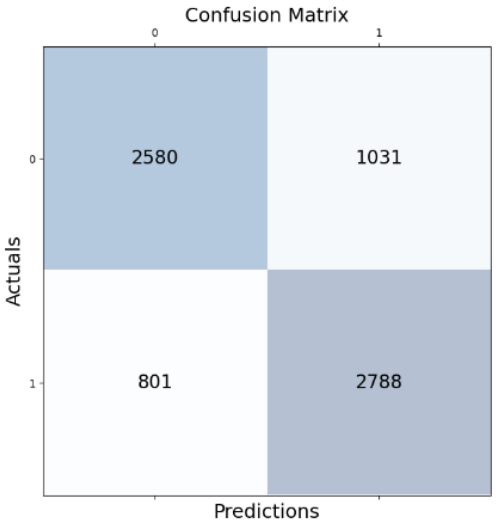
- Vstupné parametre(16) - score,D_owners,D_release_date,is_free,languages,publisher_est,self_published,has_dlc,has_website_linked,has_controller_support,is_multi_player,is_early_access,Indie,Action,Casual,Adventure,Singleplayer
- Použitie len Top 5 žánrov namiesto 9
- 1 skrytá vrstva s 16 neurónmi a aktivačnou funkciou relu

- Počet epoch = 1000
- EarlyStopping s patience=100



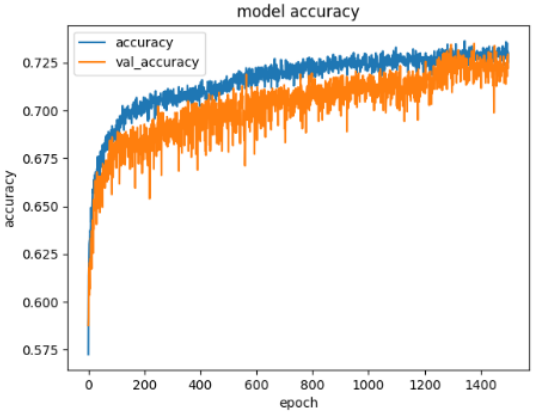

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.74 | 0.65 | 0.69 | 500 |
| 1.0 | 0.69 | 0.77 | 0.73 | 500 |
| accuracy | | | 0.71 | 1000 |
| macro avg | 0.71 | 0.71 | 0.71 | 1000 |
| weighted avg | 0.71 | 0.71 | 0.71 | 1000 |

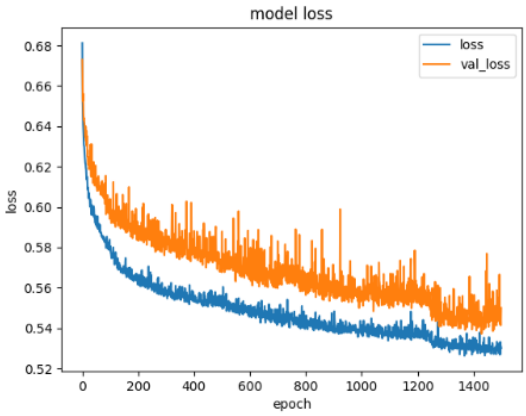
- Trénovanie ukončené po 1000. epoche.
- Úspešnosť na trénovacích dátach je 74% a na validačných 72%
- Úspešnosť na testovacích dátach je 71%

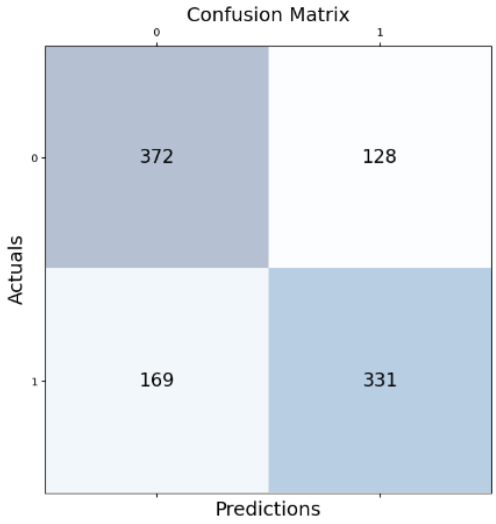
## Sieť 4

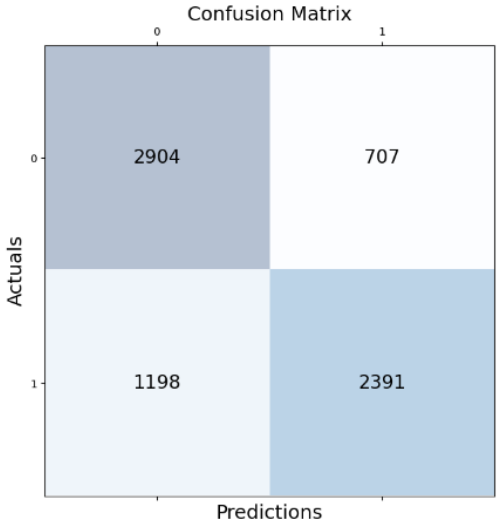
- Vstupné parametre(14) - score,D_owners,D_release_date,is_free,languages,publisher_est,self_published,has_dlc,has_website_linked,is_multi_player,Indie,Action,Casual,Adventure,Singleplayer
- Odstránili sme is_early_access a has_controller_support

- 1 skrytá vrstva s 14 neurónmi a aktivačnou funkciou relu
- Počet epoch = 1500
- EarlyStopping s patience=100




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.69 | 0.74 | 0.71 | 500 |
| 1.0 | 0.72 | 0.66 | 0.69 | 500 |
| accuracy | | | 0.70 | 1000 |
| macro avg | 0.70 | 0.70 | 0.70 | 1000 |
| weighted avg | 0.70 | 0.70 | 0.70 | 1000 |

- Trénovanie ukončné po 1500. epoche.
- Úspešnosť na trénovacích dátach je 73% a na validačných 72%.
- Úspešnosť na testovacích dátach je 70%.

## Záver

Dosiahli sme výsledky v rozsahu 68-71%. Pričom najlepšia sieť bola sieť 3. Mala 16 vstupných parametrov a 1 skrytú vrstvu so 16 neurónmi. Pri prvej sieti bol rozdiel medzi trénovacími

a testovácimi dátami 5%. Po pridaní early-stoppingu v druhej sieti sme dosiahli rozdiel 2%, ktorý bol najmenší zo všetkých sietí.