Zadanie 1 - Neurónové siete

Pavol Polednák

Contents

1 Predstavenie

Cieľom tohto projektu je analyzovať a predpovedať emocionálny vplyv piesní na základe ich hudobných vlastností. Hudba vždy zohrávala kľúčovú úlohu pri vyjadrovaní a vyvolávaní emócií a s príchodom digitálnych hudobných platforiem je k dispozícii množstvo údajov, ktoré popisujú rôzne vlastnosti skladieb. Využitím týchto údajov sa snažíme pochopiť vzťah medzi týmito hudobnými prvkami a emóciami, ktoré vyvolávajú. Súbor údajov použitý v tomto projekte obsahuje skladby zo Spotify, z ktorých každá sa vyznačuje niekoľkými vlastnosťami, ako je tanečnosť, energia, tempo a ďalšie. Každá skladba je navyše označená emóciou, čo poskytuje jasný cieľ pre naše prediktívne modelovanie.

Riešenie bude v jazyku Python. Údaje sú vo forme .csv súboru, z ktorého bude vytvorený DataFrame objekt knižnice pandas. Riadky tohoto DataFrame objektu budeme nazývať vzorky(observations/samples) a jeho stĺpce budeme nazývať vlastnosti(features).

Hlavnými bodmi tohto projektu sú:

1. **Čistenie a predbežné spracovanie údajov**: Zabezpečenie toho, aby bol súbor údajov čistý, bez extrémnych hodnôt a vhodne formátovaný na analýzu.
2. **Analýza prieskumných údajov (EDA)**: Vykonávanie dôkladnej analýzy s cieľom pochopiť distribúciu, trendy a vzorce údajov.
3. **Prediktívne modelovanie**: Vytvorenie modelu strojového učenia na predpovedanie emócií skladby na základe jej vlastností.

Dokumentácia poskytuje prehľad použitých metodík, zistení EDA a výkonnosti prediktívneho modelu. Slúži ako sprievodca pre pochopenie projektu a poznatkov, ktoré z neho vyplývajú.

2 Úprava vlastností

2.1 Predstavenie

Dáta tvoria záznamy piesní a ich vlastností.

- **Vlastnosti s normalizovaným rozsahom <0-1>**:
  - *Tanečnosť (Danceability)*: Kvantifikuje vhodnosť skladby na tanec.
  - *Valencia (Valence)*: Reprezentuje hudobnú pozitívnosť, ktorú skladba sprostredkováva.
  - *Energia (Energy)*: Meria intenzitu a dynamický rozsah skladby.
  - *Živosť (Liveness)*: Určuje pravdepodobnosť, že skladba je živou nahrávkou.
  - *Rečovosť (Speechiness)*: Rozlišuje medzi rečovými skladbami a hudbou.
  - *Nástrojovosť (Instrumentalness)*: Predpovedá absenciu vokálneho obsahu.
  - *Akustickosť (Acousticness)*: Meria množstvo akustického zvuku v skladbe.
- **Vlastnosti súvisiace s žánrom**:
  - *Žánre (Genres)*: Zoznam žánrov spojených so skladbou.
  - *Filtrované žánre (Filtered_ genres)*: Žáner po filtrácii na základe dostupných parametrov.
  - *Hlavný žáner (Top_ genre)*: Najvýraznejší žáner spojený so skladbou.
- **Technické zvukové vlastnosti**:
  - *Hlasitosť (Loudness)*: Celková hlasitosť v decibeloch (dB), ktorá je logaritmickou jednotkou.
  - *Tempo*: Rýchlosť skladby meraná v BPM (údery za minútu).
  - *Dĺžka v ms (Duration_ ms)*: Celková dĺžka skladby vyjadrená v milisekundách.
- **Kategorické a binárne vlastnosti**:
  - *Explicitný obsah (Explicit)*: Binárna vlastnosť označujúci prítomnosť explicitného obsahu.
  - *Počet umelcov (Number_ of_ artists)*: Udáva počet umelcov spojených so skladbou.
- **Všeobecné vlastnosti**:
  - *Popularita (Popularity)*: Číselné skóre reprezentujúce popularitu skladby.
  - *Názov (Name)*: Názov skladby.
  - *URL adresa (URL)*: Webový odkaz na skladbu.

Ďalej sa budú vyuzívať anglické názvy vlastností.

Časťou úlohy bolo zamyslieť sa nad odstránením niektorých stĺpcov z datasetu, a to takých, ktoré nebudú v neskorších častiach použité. Rozhodli sme sa, že tieto budú:

- Žánre: Táto vlastnosť obsahuje množinu žánrov vyhovujúcich danej piesni. Po zakódovaní všetkých unikátnych žánrov One-hot encodingom vznikol DataFrame s 1490 stĺpcami. Rozhodli sme sa preto túto vlastnosť vynechať a riadiť sa podľa hlavného žánra.

- Filtrované žánre

- Názov/URL adresa: Duplikáty boli z dát odstránené. Dá sa predpokladať, že existujú unikátne URL a názvy piesní a práca s nimi by nepriniesla závery, ktoré sú hodné zamyslenia.

Pre tieto vlastnosti nebude zostavená podsekcia, ani sa neobjavia pri trénovaní prvého modelu.

Pre ostatné vlastnosti sa tvoril histogram aj boxplot. Pre zistenie existencie outlierov poskytneme box plot pred úpravou dát a pre zobrazenie distribúcie po úprave poskytneme histogram.

Pre vlastnosti s kontinuálnymi číselnými hodnotami budeme taktiež po úprave dát sledovať tri hodnoty:

- *Priemer:* Priemerná hodnota množiny čísel vypočítaná sčítaním všetkých čísel a následným vydelením počtom týchto čísel. :

- *Medián:* Stredná hodnota v množine čísel, ak sú usporiadané v poradí; Ak existuje párny počet pozorovaní, je to priemer dvoch stredných čísel.

- *Smerodajná odchýlka:* Miera veľkosti odchýlky alebo rozptylu v množine hodnôt, ktorá udáva, ako sú čísla rozložené od priemeru. Nízka smerodajná odchýlka znamená, že hodnoty majú tendenciu byť blízke strednej hodnote, zatiaľ čo vysoká smerodajná odchýlka znamená, že hodnoty sú rozložené v širšom rozsahu.

2.2 Metodológia úpravy dát

Rozhodli sme sa pre rôznorodé postupy úpravy dát, podľa kontokrétnych vlastností. Z dôvodu veľmi vysokého počtu vzoriek sme posúdili, že bude vhodnejšie vzorky s chýbajúcimi, či nesmyselnými hodnotami úplne odstrániť.

- Pre vlastnosti ohraničené <0, 1> sme odstránili akékoľvek záznamy mimo tejto hranice.
- Pre vlastnosť popularita sme odstránili všetky záznamy mimo určenej hranice <0, 100>.
- Pre tempo a dĺžku sme odstránili nezmyselné(nulové a záporné) hodnoty. V prípade dĺžky sme sa taktiež rozhodli odstániť vzorky, ktoré boli príliš krátke pre účely piesne. Ako zhovievanú hranicu sme si zvolili 40 sekúnd.
- Pre hlasitosť, tempo a dĺžku sme sa rozhodli odstrániť hodnoty príliš vzdialené od stredných, použili sme na to to Interquartile Range (IQR) metódu. Použili sme koeficienty 1.5 pre hlasitosť a tempo, 2.5 pre dĺžku (ktorá mala veľké množstvo outlierov).
- Vlastnosti, v ktorých ešte neprebehla kontrola proti nulovým hodnotám sú: počet umelcov, explicitný obsah, hlavný žáner, emócie.
- Ako poistku taktiež odstránime nebinárne hodnoty z vlastnosti explicitný obsah, odstránime počty umelcov menšie alebo rovné 0. Zároveň odstránime duplikáty.

Pôvodný DataFrame mal rozmery: (11960, 19)

Metóda úpravy	Počet odstránených vzoriek
Odstránenie duplikátov	2
Odstránenie vzoriek vlastností mimo hraníc <0,1>	16
Odstránenie vzoriek pre tempo, duration a popularity	175
Filtrácia outlierov s IQR metódou	498
Odstránenie vzoriek s chýbajúcimi hodnotoami	277
Sanity check pre zvyšné vzorky	0

Table 1: Úprava dát

Po úprave dát maá DataFrame rozmery: (10992, 15)

2.3 Univariátna analýza

2.3.1 Danceability

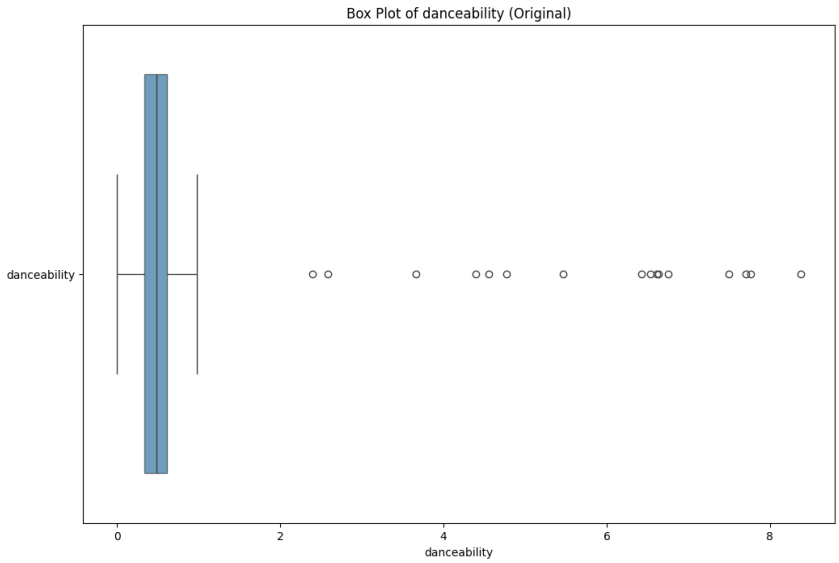

Figure 1: Box plot pre danceability - pôvodný

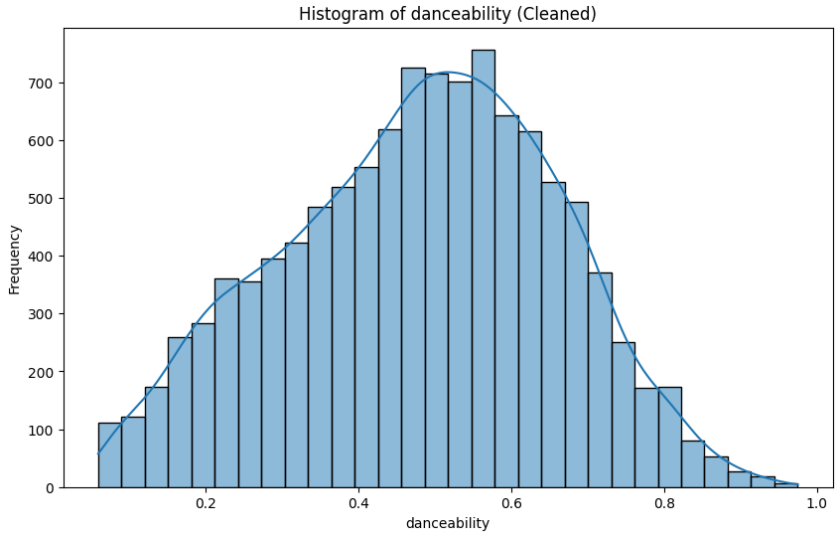

Figure 2: Histogram pre danceability - po úprave

Vlastnosti meria, aká vhodná je skladba na tanec na základe tempa, stability rytmu, sily rytmu a celkovej pravidelnosti. Z histogramu 2 možeme vidieť, distribúciu týchto hodnôt. Pretiahnutie box plotu 1 až k hodnotám okolo 9 značí, že existovali neželané outliary >1. Úpravou dát sme tieto outliary odstránili.

- Priemer: 0.4791
- Medián: 0.4910
- Štandardná odchýlka: 0.1738

2.3.2 Valence

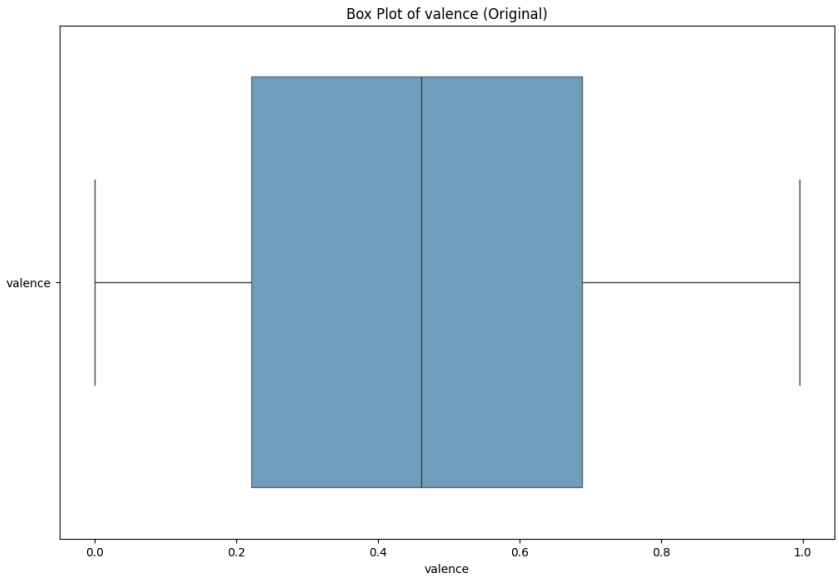

Figure 3: Box plot pre valence - pôvodný

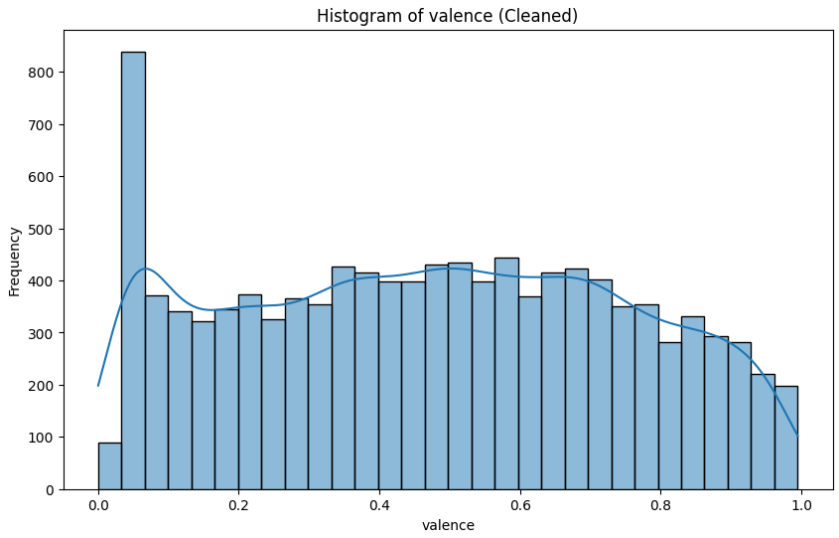

Figure 4: Histogram pre valence - po úprave

Vlastnosť predstavuje hudobnú pozitívnosť skladby. Vyššie hodnoty naznačujú pozitívnejšie nálady. Outliery sa vo vlastnosi nenachádzali.

- Priemer: 0.4715
- Medián: 0.4750
- Štandardná odchýlka: 0.2729

2.3.3 Energy

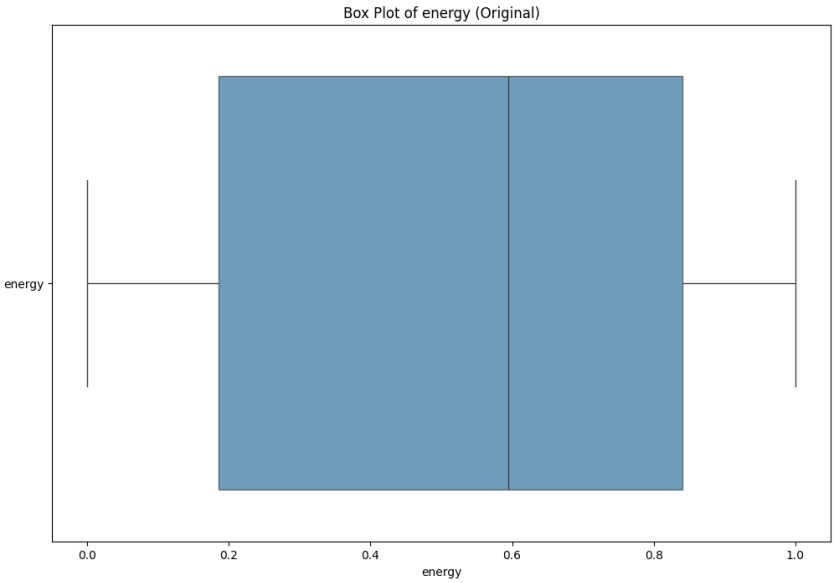

Figure 5: Box plot pre energy - pôvodný

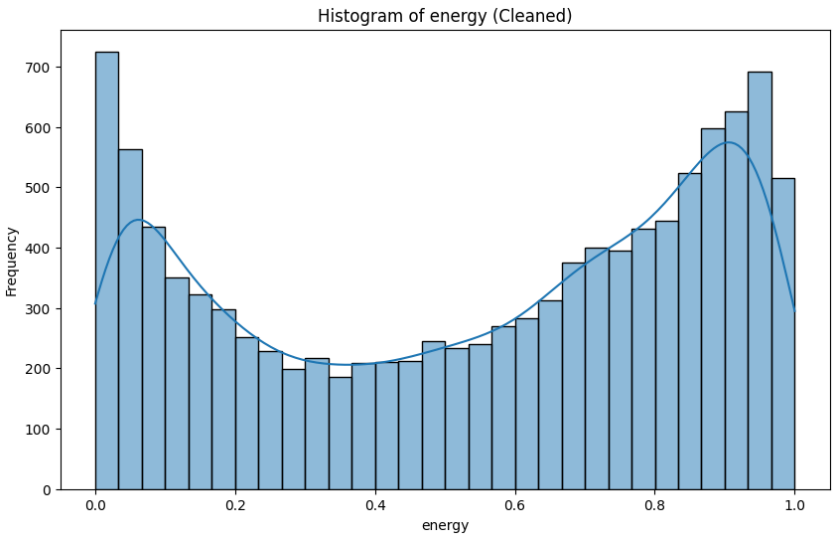

Figure 6: Histogram pre energy - po úprave

Vlastnosť meria intenzitu a aktivitu. Energické skladby sú zvyčajne rýchle, hlasné a hlučné. Podľa histogramu vidíme, že väčšina piesní sa silno prikláňala na jednu stranu spektra. Outliery sa vo vlastnosti nenachádzali.

- Priemer: 0.5407
- Medián: 0.6145
- Štandardná odchýlka: 0.3294

2.3.4 Liveness

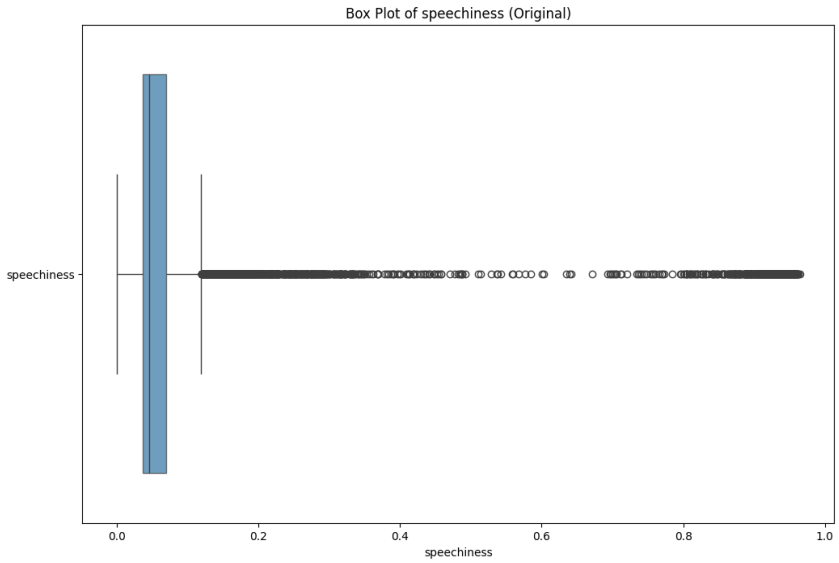

Figure 7: Box plot pre liveness - pôvodný

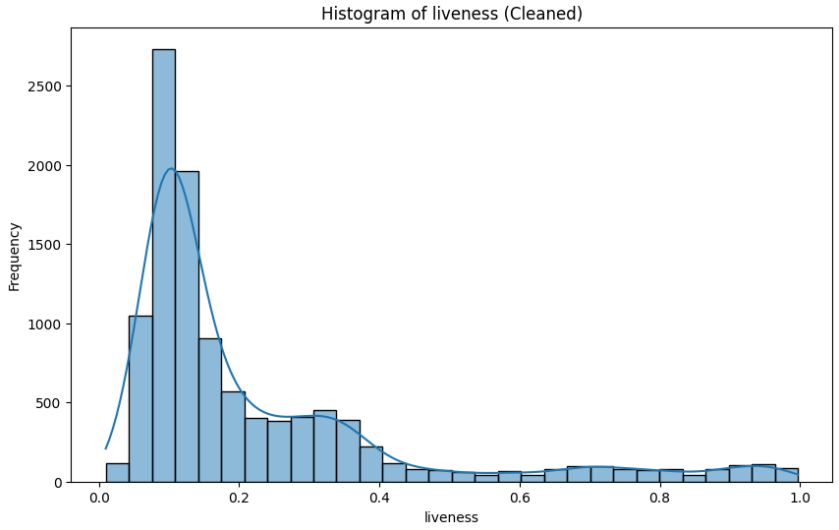

Figure 8: Histogram pre liveness - po úprave

Vlastnosť zistí prítomnosť publika v nahrávke. Vyššie hodnoty predstavujú vyššiu pravdepodobnosť, že skladba bude vykonaná naživo. Podľa očakávaní, táto vlastnosť nezvykla mať vysoké hodnoty. Outliery sa vo vlastnosti nenachádzali.

- Priemer: 0.2260
- Medián: 0.1315
- Štandardná odchýlka: 0.2128

2.3.5 Speechiness

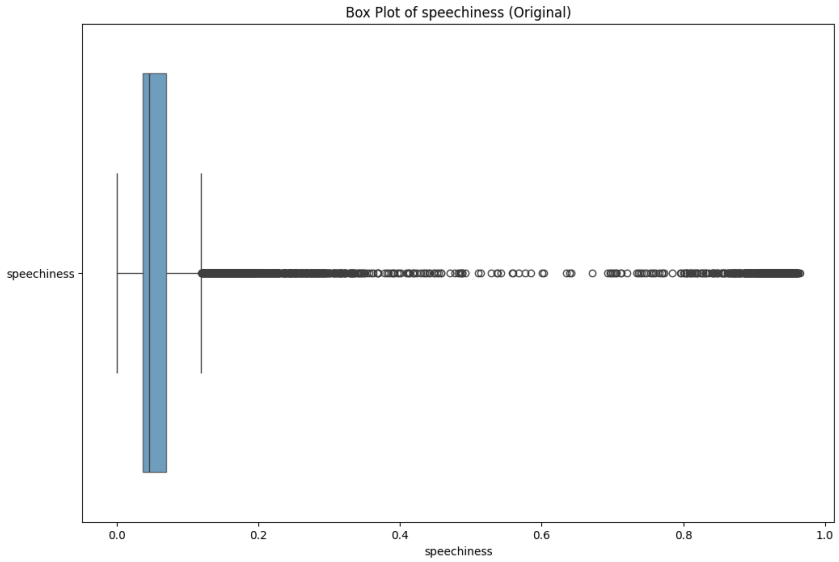

Figure 9: Box plot pre speechiness - pôvodný

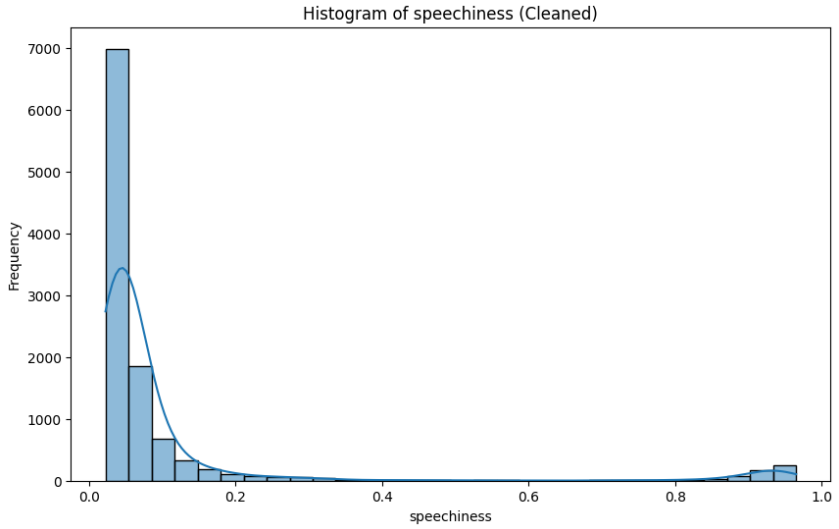

Figure 10: Histogram pre speechiness - po úprave

Vlastnosť meria prítomnosť hovoreného slova v stope. Skladby s vyššou rečou sú zhovorčivejšie. Stála nízka úroveň týchto hodnôt predstavuje pomerne veľké prekvapenie. Outliery sa vo vlastnosti nenachádzali.

- Priemer: 0.1049
- Medián: 0.0459
- Štandardná odchýlka: 0.1932

2.3.6 Instrumentalness

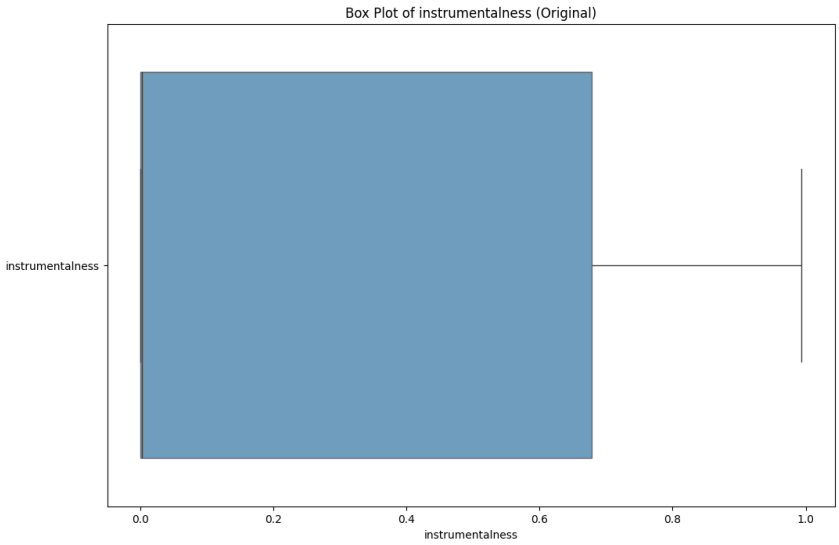

Figure 11: Box plot pre instrumentalness - pôvodný

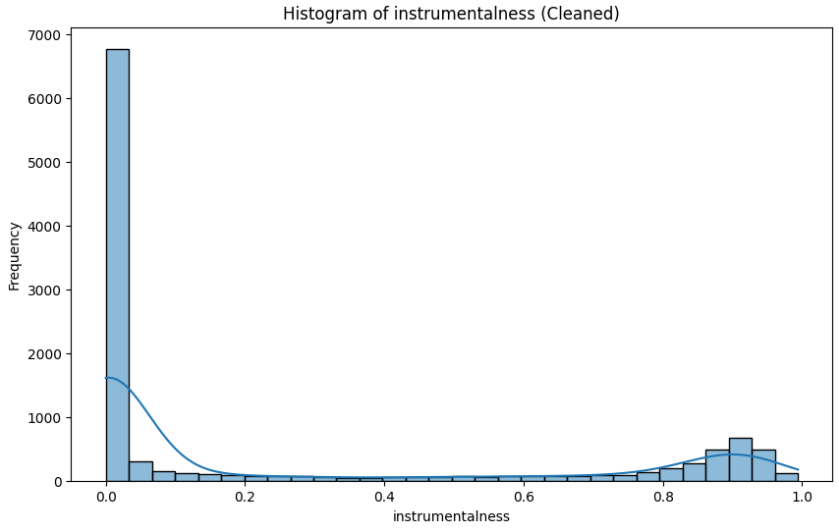

Figure 12: Histogram pre instrumentalness - po úprave

Vlastnosť predpovedá, či skladba obsahuje vokály. Vyššie hodnoty naznačujú, že skladba je inštrumentálna. Viac ako polovica piesní bola úplne vokálna, avšak vyskytlo sa aj niekoľko zameraných na hudobné nástroje. Outlduriery sa vo vlastnosti nenachádzali.

- Priemer: 0.2493
- Medián: 0.0018
- Štandardná odchýlka: 0.3712

2.3.7 Acousticness

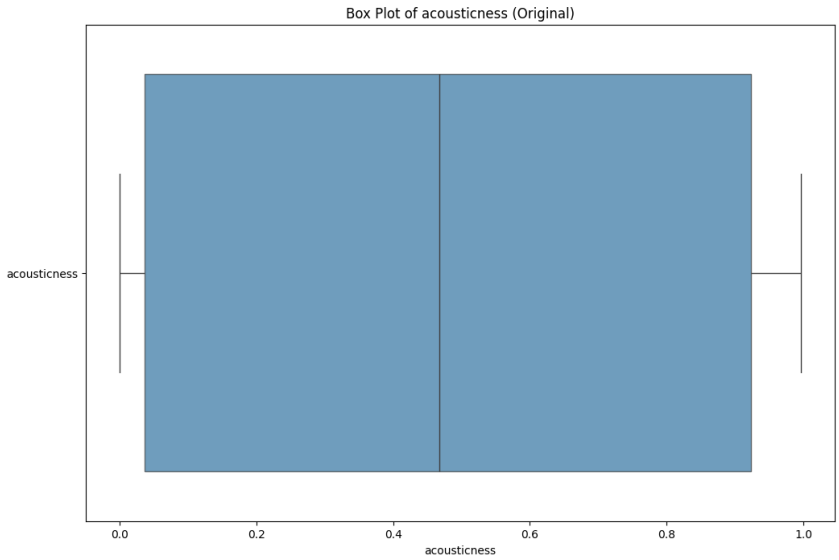

Figure 13: Box plot pre acousticness - pôvodný

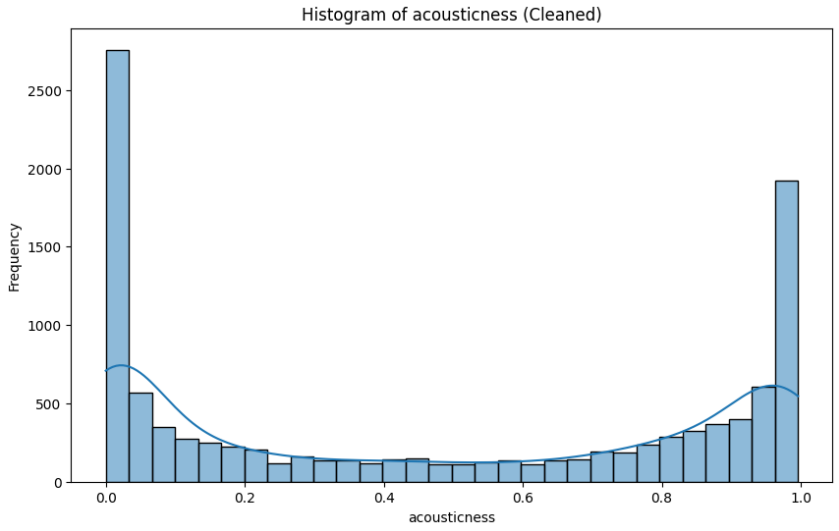

Figure 14: Histogram pre acousticness - po úprave

Vlastnosť predstavuje množstvo akustického zvuku v stope. Vyššie hodnoty znamenajú viac akustických zvukov. Svojím tvarom histogram pripomína viac extrémnu verziu histogramu energie. Outliery sa vo vlastnosi nenachádzali.

- Priemer: 0.4735
- Medián: 0.4420
- Štandardná odchýlka: 0.4035

2.3.8 Loudness

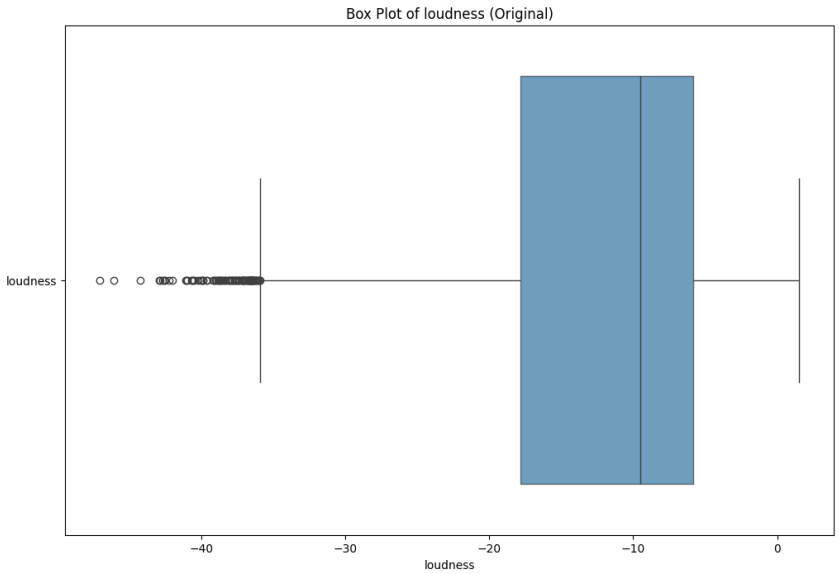

Figure 15: Box plot pre loudness - pôvodný

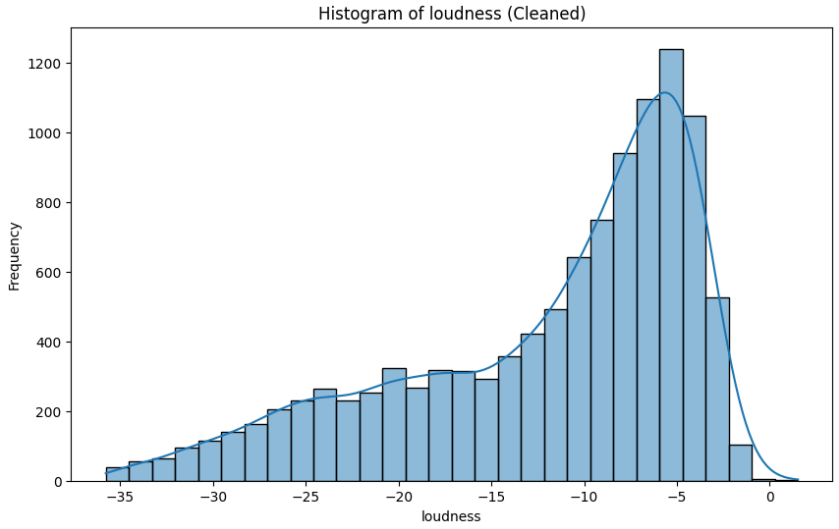

Figure 16: Histogram pre loudness - po úprave

Vlastnosť meria celkovú hlasitosť stopy v decibeloch (dB). Hodnoty hlasitosti tratí sú spriemerované na celej trati. Histogram ukazuje jasnú preferenciu pre hlasnejšie skladby. Outliery sa vo vlastnosti nachádzali, avšak len ojedinele.

- Priemer: -12.0012
- Medián: -9.2735
- Štandardná odchýlka: 7.9116

2.3.9 Tempo

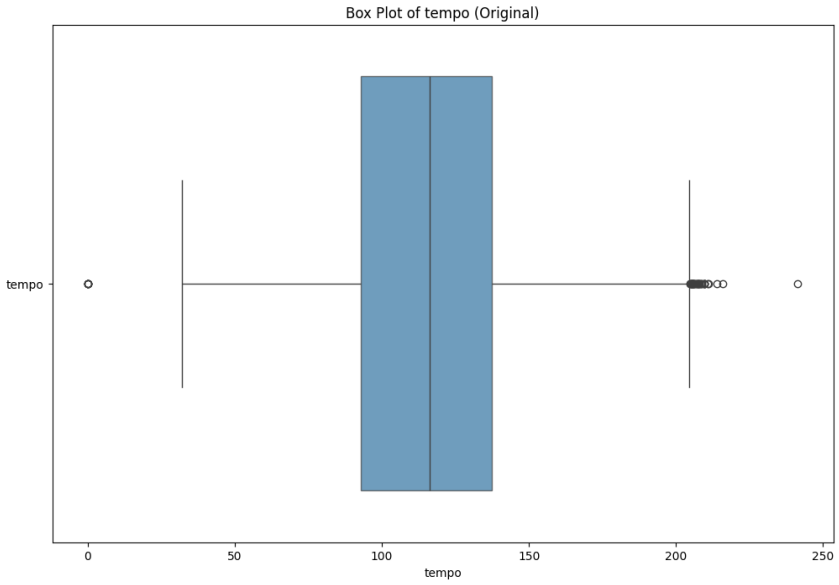

Figure 17: Box plot pre tempo - pôvodný

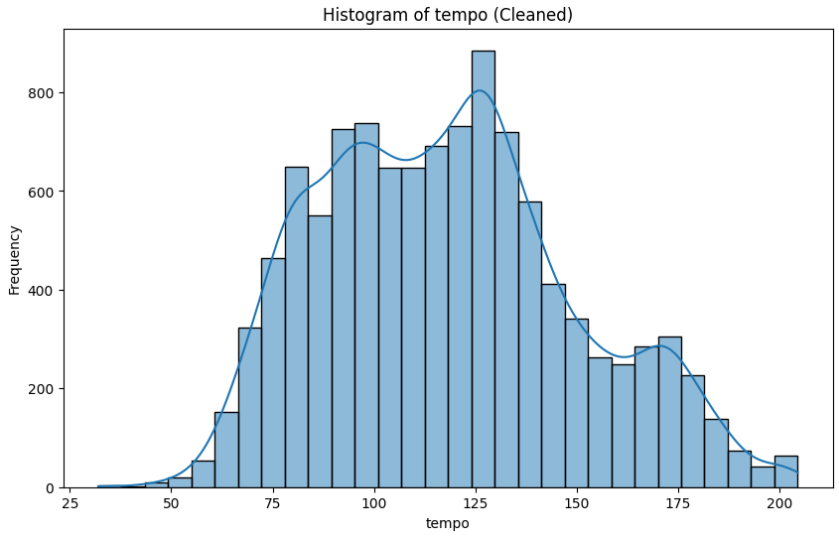

Figure 18: Histogram pre tempo - po úprave

Vlastnosť predstavujeýchlosť alebo tempo danej skladby, merané v BPM (Beats Per Minute). Outliery sa vo vlastnosti nachádzali, a to hodnoty ako nereálne vysoké, či nulové tempo.

- Priemer: 118.1759
- Medián: 116.8070
- Štandardná odchýlka: 31.4584

2.3.10 Duration in ms

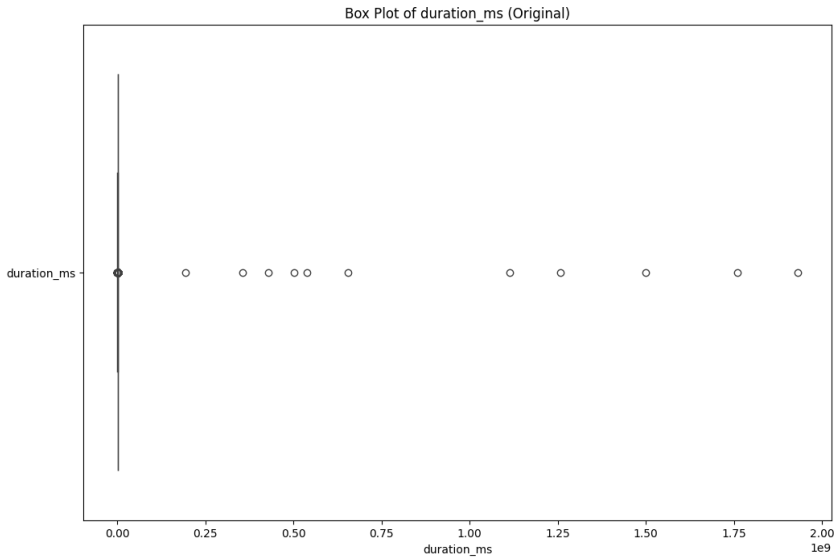

Figure 19: Box plot pre duration - pôvodný

Trvanie dráhy v milisekundách. Histogram naznačuje outliery v podobe neskutočne dlhých piesní. Taktiež existovali viacero piesní s nulovou alebo negatívnou dĺžkou. Pridávame histogram po použití Interquartile Range (IQR) metódy na filtráciu outlierov z dát.

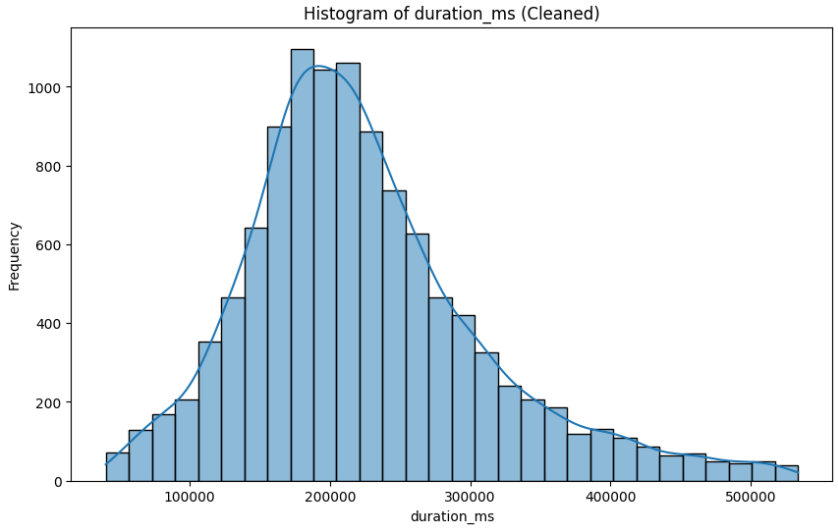

Figure 20: Histogram pre duration - upravený

- Priemer: 224,429s
- Medián: 211,299s
- Štandardná odchýlka: 86,288s

2.3.11 Popularity

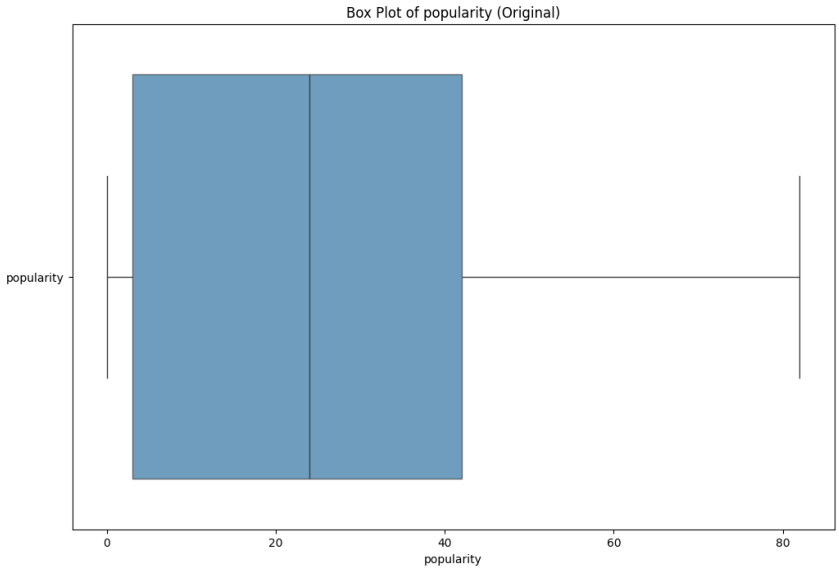

Figure 21: Box plot pre popularity - pôvodný

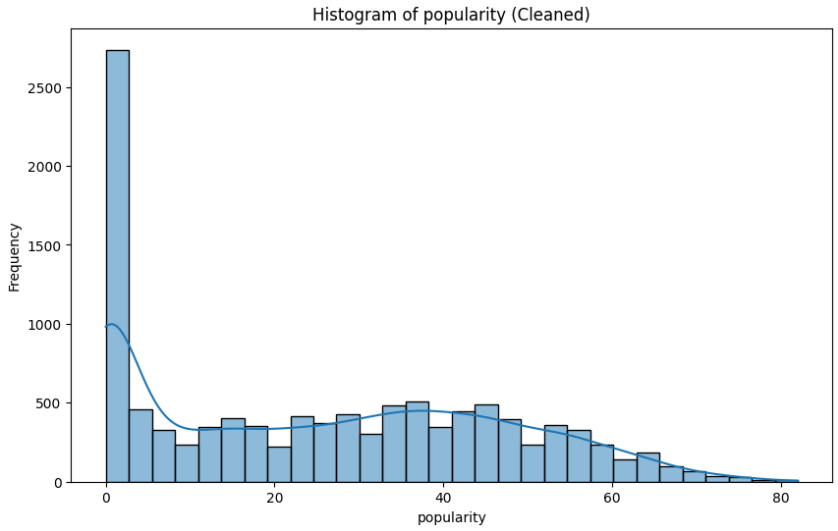

Figure 22: Histogram pre popularity - po úprave

Meria popularitu skladby. Hodnota je založená na celkovom počte prehratí. Hodnota by mala byť v rámci <0,100>. Zrejme nikoho neprekvapí, že patrónka umení sa neusmeje na každého.

- Priemer: 25.3670
- Medián: 25.0000
- Štandardná odchýlka: 21.1509

2.3.12 Explicit content

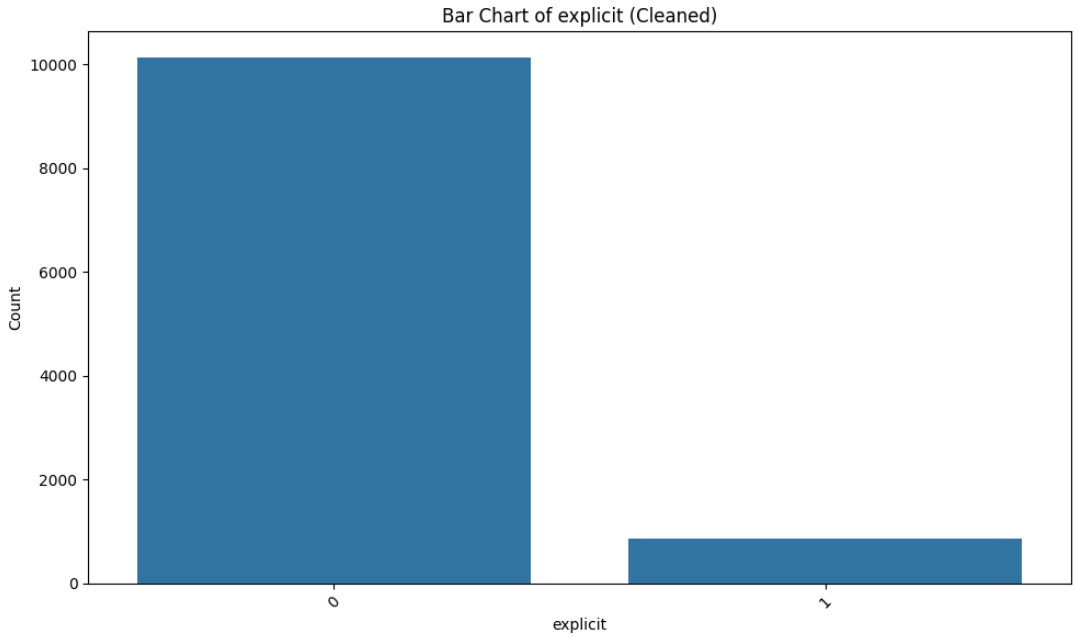

Figure 23: Stĺpcový graf pre explic. obsah - po úprave

Označuje, či má stopa explicitný obsah alebo nie, ktorý predstavuje binárnu hodnotu. Outliery pre túto vlastnosť neexistovali.

2.3.13 Počet umelcov

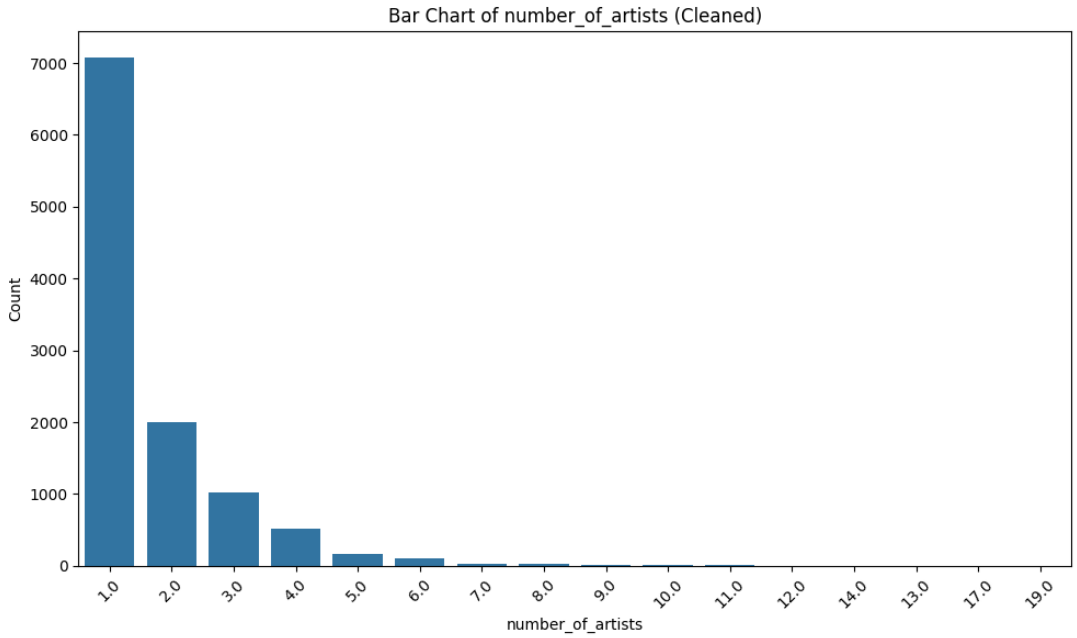

Figure 24: Stĺpcový graf pre počet umelcov - po úprave

Meria, koľko umelcov sa zapojilo do tvorby piesne Outliery pre túto vlastnosť neexistovali. Všetky hodnoty boli nad 0 a rozhodli sme sa nechať aj maximálne hodnoty v tejto vlastnosti - 19.

2.3.14 Hlavný žáner

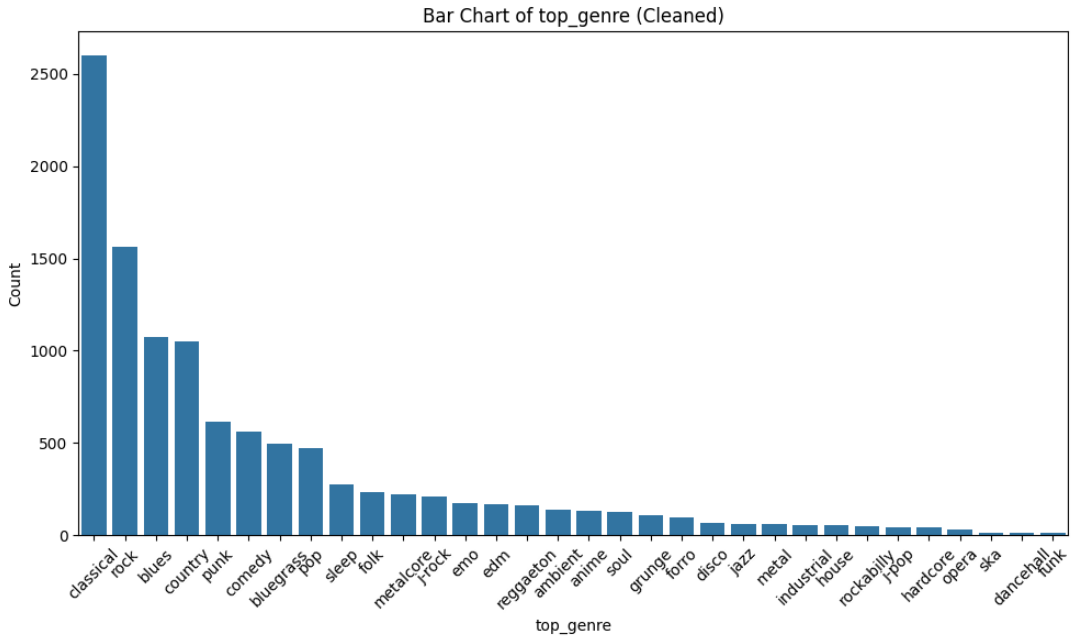

Figure 25: Stĺpcový graf pre hlavný žáner - po úprave

Označuje najikonickejší žáner pre tvorcov piesne. Outliery pre túto vlastnosť neexistovali.

2.3.15 Emócie

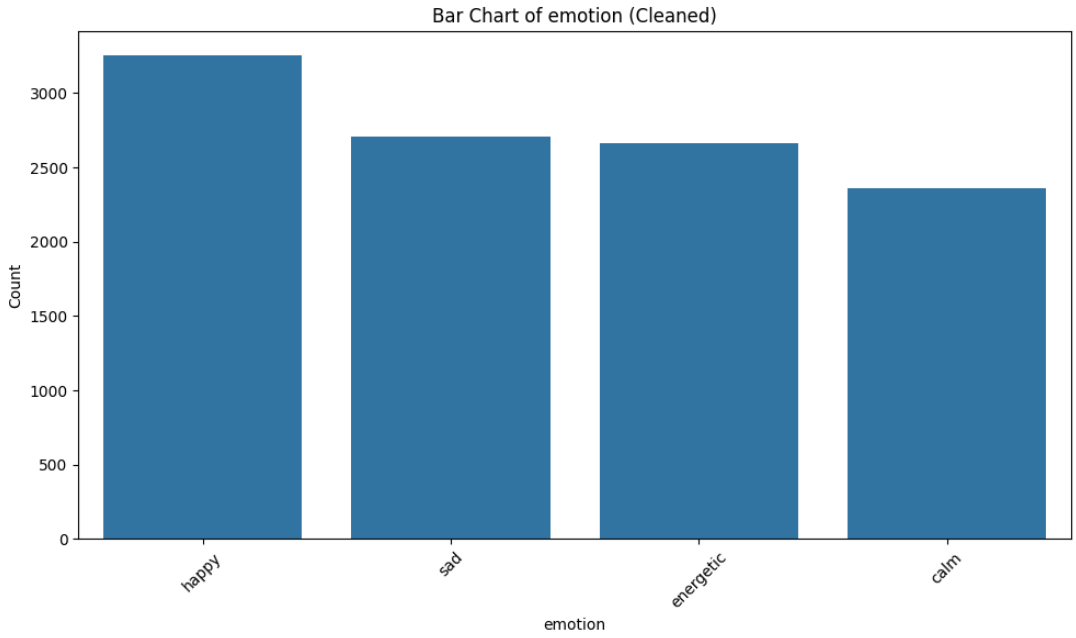

Figure 26: Stĺpcový graf pre emócie - po úprave

Označuje, aká emócia je najsilnejšie spojená so skladbou. Je to náša výstupná vlastnosť. Outliery pre túto vlastnosť neexistovali.

3 EDA

EDA - Exploratory Data Analysis. Pokiaľ ide o strojové učenie, EDA je kritickým krokom, ktorý zahŕňa analýzu a skúmanie údajov a zhrnutie ich hlavných charakteristík, pričom často využíva metódy vizualizácie.

Univariátnou(jednorozmernou) analýzou sme prešli. Je to analýza jednej vlastnosti a jej cieľom je popis dát a nájdenie vzorov(patterns). Časťou zadanie je nájdenie piatich zaujímavých vzťahov v dátach. Pre tento cieľ využijeme bivariátnu(dvojrozmernú) analýzu. Jej cieľom je súbežné skúmanie dvoch vlastností, s cieľom zistenia vzťahu medzi nimi.

Na úvod prestavíme niekoľko hypotéz a ichsprávnosť skúsime overiť.
Hyp. 1: Piesne s vyšším tempo budú vhodnejšie pre tanec.
Hyp. 2: Piesne s väčšou energiou budú hlasnejšie.
Hyp. 3: Piesne s vyššou mierou násjtrovosti budú mať vyššiu mieru akustickosti.
Hyp. 4: Piesne s viac pozitívnymi emóciami bude vhodnejšia na tanec.
Hyp. 5: Piesne s nižšou hlasitosťou budú viac akustickejšie.

3.1 Hypotéza 1

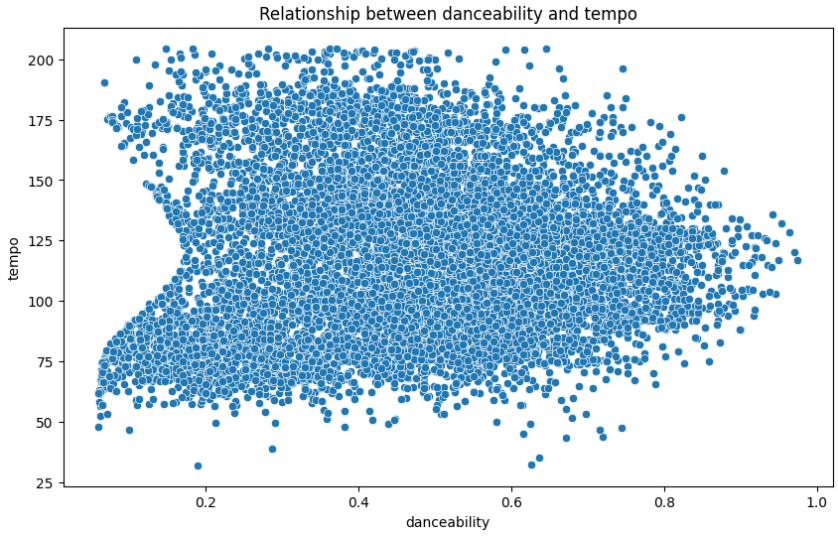

Figure 27: Scatter plot pre danceability a tempo

Danceability má široké spektrum naprieč celým tempom a teda samotné tempo nebude významné pre predpoveď danceability. Existujú mnohé vzorky s vysokým tempom aj danceability, avšak mnohé majú vysoké tempo a nízku danceability. Hypotéza podklad nemá.

3.2 Hypotéza 2

Scatter plot opisuje pozitívny trend, vyššia energie je asociovaná s vyššou hlasitosťou. Toto je indikované koncentráciou bodov z ľavého spodného rohu(nízka energia, nízka

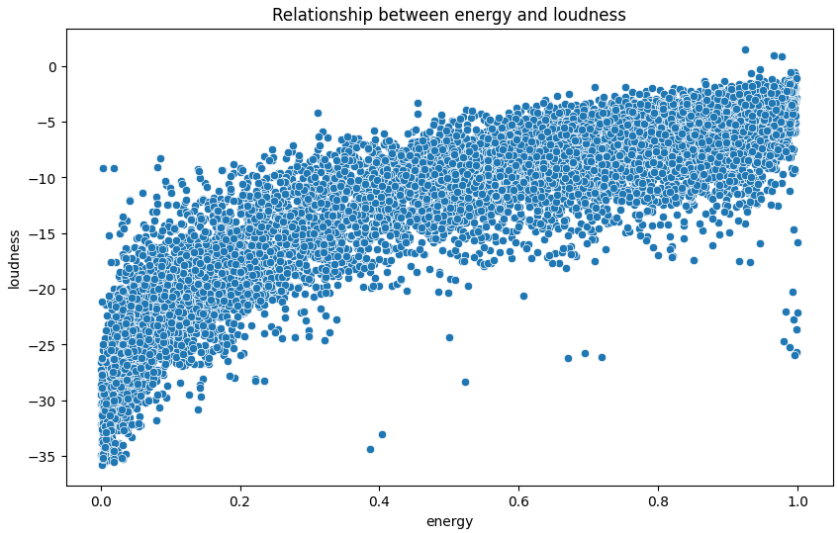

Figure 28: Scater plot pre energy a loudness

hlasitosť), do pravého horného rohu(vysoká energia, vysoká hlasitosť). Hypotéza je podložená, teda máme **vzťah č. 1**

3.3 Hypotéza 3

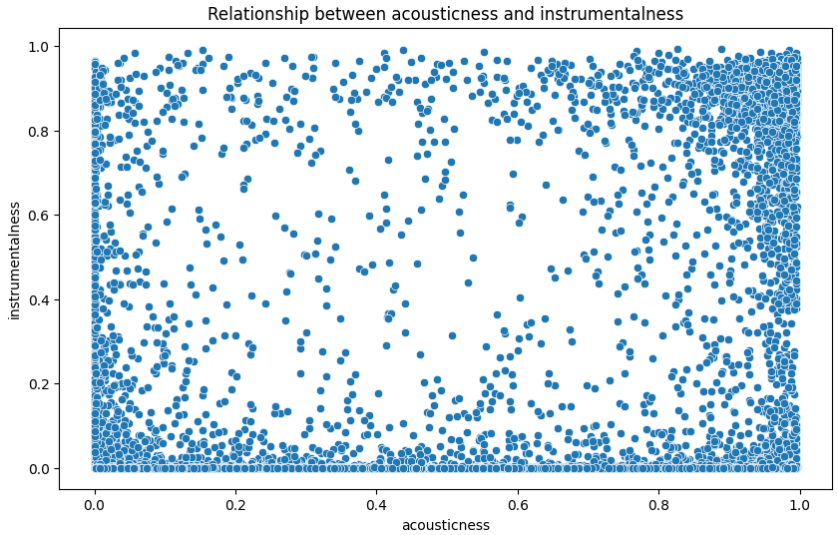

Figure 29: Stĺpcový graf pre počet umelcov - po úprave

Scatter plot ukazuje, že piesne s vysokou mierou nástrojovosti majú rôznu mieru akustickosti, vrátane mnohých s nízkou akustickosťou. Aj napriek zaujímavému tvaru výsledného grafu teda nemôžeme povedať, že by existoval viac ako mierny pozitívny vzťah medzi vlastnosťami.

3.4 Hypotéza 4

Napriek tomu, že sa v grafe črtá vzor, hodnoty vzoriek sú stále pomerne široké. Môžeme však povedať, že piesne s nízkou valenciou takmer isto nebudú mať vysokú mieru tanečnosti

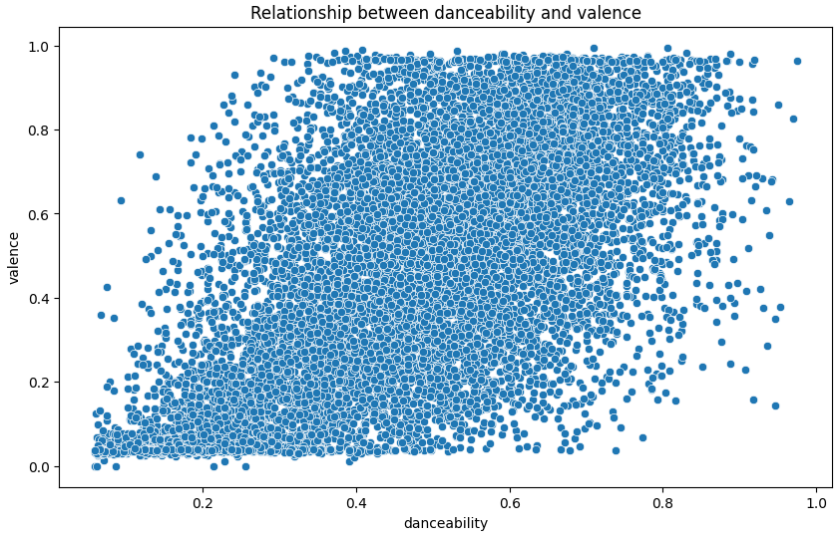

Figure 30: Scatter plot pre danceability a valence

a piesne s vysokou valenciou nebudú mať nízku mieru tanečnosti. Toto bude **vyťah č. 2**

3.5 Hypotéza 5

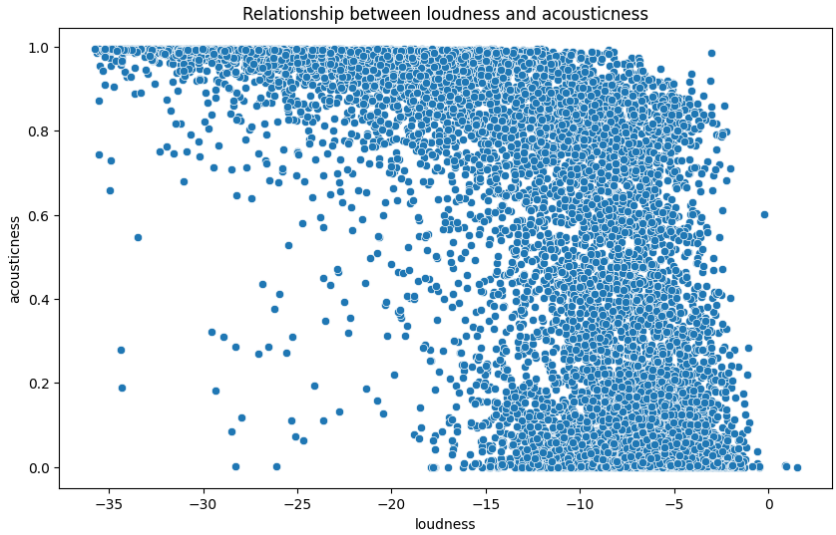

Figure 31: Scater plot pre loudness a acousticness

Scatter plot ukazuje, že existuje veľké množstvo piesní s vysokou hlasitoťou, a to vo všetkých mierach akustickosti. Toto môže mať viacero príčin, vrátane nahrávacích a produkčných techník. Graf však vytvára jasný vzor, kde piesne s veľmi nízkou hlasitosťou majú nepomerne vysokú šancu mať vysokú akustickosť. Toto bude **vzťah č. 3**

3.6 Heat map

Nie všetky vzťahy sú zistiteľné týmto spôsobom. Máme môžnosť vytvoriť heat map(teplotnú mapu) a v nej zobraziť korelačnú maticu. Táto matica zobrazuje korelačné koeficienty

medzi dvoma vlastnosťami. Každá bunka v tabuľke bude predstavovať koreláciu medzi dvomi z nich.

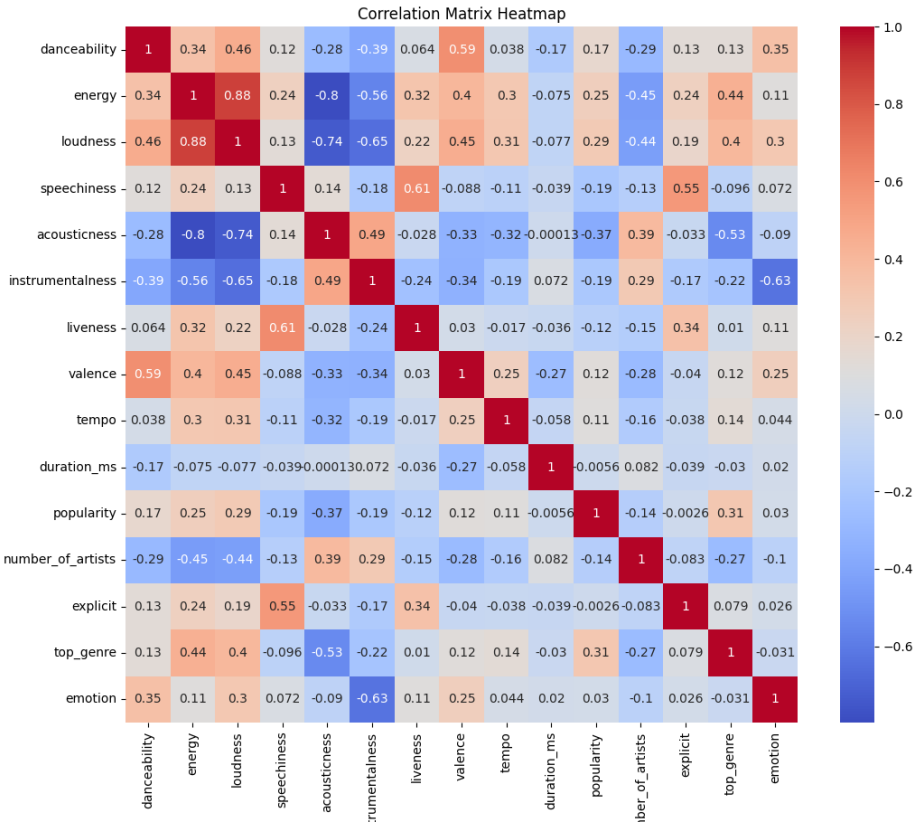

Figure 32: Heat map pre DataFame

Výrazne pozitívne čísla(odtiene červenej) označujú pozitívnu koreláciu, zatiaľ čo výrazne negatívne(odtiene modrej) negatívnu koreláciu. Pre hľadanie vzťahov medzi vlastnosťami sa snažíme hľadať extrémne hodnoty.

Ako posledné dva vzťahy si sme vybrali vzťahy valencie a inštrumentality voči emócii piesne.

Graf ukazuje zaujímavú tendenciou, kde piesne s vysokou mierou použitia hudobných nástrojov patrí zpravidla do kategórie pokojných piesní. Existujú výnimky, avšak toto sa môže javiť nečakane. Možným dôvodom je tenká hranica medzi zaradením do kategórií pokojných/smutných piesní do správnych kategórii. **Vzťah č.4**

Tento vzťah je zrejme viac očakávaný. Piesne označené ako šťastné majú tendenciou mať vysokú valenciu, energetické piesne sú zmiešané vrece, zatiaľ čo smutné a pokojné piesne majú tendenciu mať valenciu nižšiu. **Vzťah č.5**

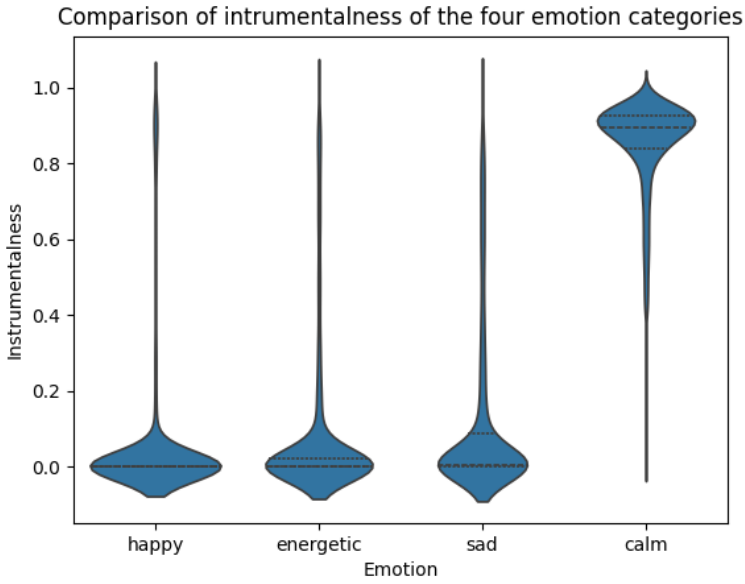

Figure 33: Violin plot vzťahu instrumentalness a emotion

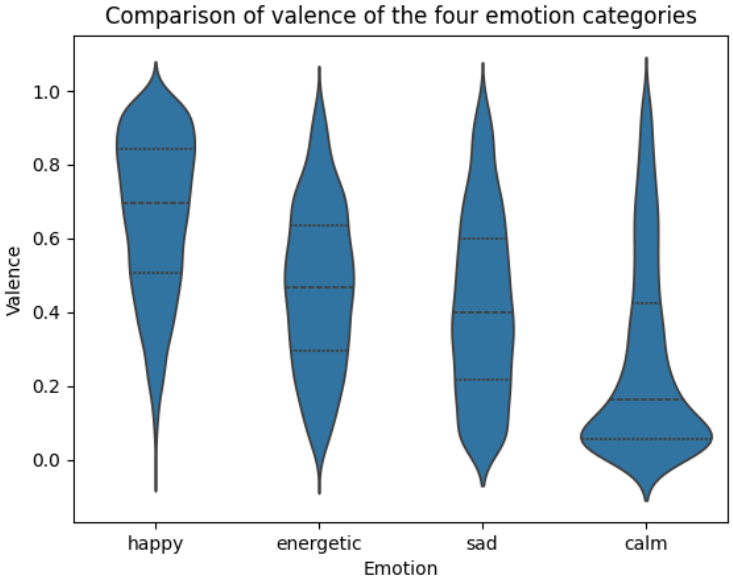

Figure 34: Violin plot vzťahu valence a emotion

4 Jednoduchá neurónová sieť

Keďže na túto časť zadania používame sklearn, na encoding sme použili metódu LabelEncoding, a to konkrétne na vlastnosti emotion a top_genre. Dáta boli rozdelené na trénovaciu/validačnú/testovaciu množinu v pomere 8:1:1 a tieto množiny boli štandardizované:

```
X_train, X_val_test, y_train, y_val_test = train_test_split(X
    , y, test_size=0.2, random_state=42)
X_valid, X_test, y_valid, y_test = train_test_split(
    X_val_test, y_val_test, test_size=0.5, random_state=42)

scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_valid_scaled = scaler.transform(X_valid)
X_test_scaled = scaler.transform(X_test)
```

Teraz môžeme prejsť na trénovanie modelu. Pre prvý pokus o jdenoduchú neurónovú sieť sme sa rozhodli pre čo najminimalistickejšie nakonfigurovaný model. Sto neurónov v jednej vrstve poskytne dostatok komplexity pre modelovanie nelineárnych vzťahov počas štartovaních experimentov. Päťsto iterácii poskytne neurónovej sieti dostatočnú šancu naučiť sa z dát a snaží sa zabrániť pretrénovaniu nastavením hornej hranice. Nastavením parametra random_state=42 zabezpečíme, že výsledok bude repodukovateľný.

```
#Model training
mlp = MLPClassifier(hidden_layer_sizes=(100),
                    max_iter=500,
                    random_state=42)
mlp.fit(X_train_scaled, y_train)
```

MLP accuracy on train set: 0.9365404298874105
MLP accuracy on test set: 0.8818181818181818
Vysoká presnosť na trénovacej množine značí, že model sa naučil rozpoznávať dáta efektívne. Presnosť na testovacej množine je podľa očakávaní o niečo nižšia, no stále vysoká, čo je dobrým znamením. Čím je rozdiel vyšší, tým väčšie obavy nám spôsobí možné pretrénovanie.

Confusion matrices - vizuálne predstavujú výkonnosť klasifikačného modelu na tréningovej súprave a testovacej súprave. Každá matica ukazuje, ako dobre model predpovedal hlavné emócie vyvolané piesňami, pričom emócie sú rozdelené do štyroch kategórií: pokojné, energické, šťastné a smutné.
Na oboch maticiach vidíme podobné vzory, čo je dobrým znamením. Najpresnejšie sa určovali pokojné piesne. Najčastejšie boli nesprávne diagnostikované ako smutné piesne, čo intuitívne dáva zmysel. Ďalší takýto prípad je kombinácia energetické/šťastné piesne, kde sa vyskytlo viacero nesprávne určených vzoriek. Medzi piesňami týchto typov je však často len tenká hranica. Jeden prípad, ktorý by nám mohol spôsobiť obavy, sú nesprávne určenia štastných/smutných piesní. Pre tento prípad by sme vysokú mieru chybovosti nečakali. Je možné, že niektoré skladby balancujú tento kontrast pre umelecké zámery, avšak tento prípad je zaujímavý pre ďalšie skúmanie.

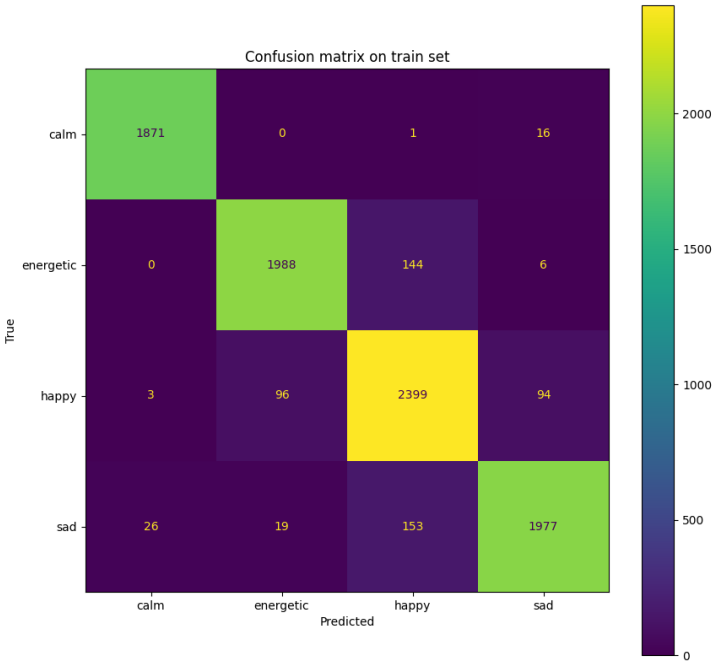

Figure 35: Confusion matrix pre trénovaciu množinu

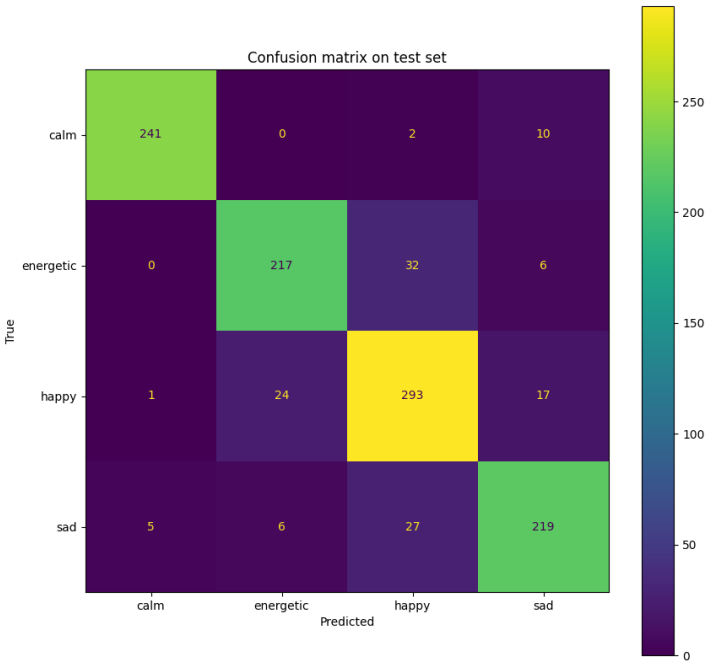

Figure 36: Confusion matrix pre testovaciu množinu

5 Neurónová sieť s použitím Keras

5.1 Pretrénovanie

Pri tejto časti sme využili OneHotEncoding namiesto LabelEncoding. Spôsobilo to nahradenie jedného stĺpca štyrmi pri enkódovaní emócie a jedného stĺpca tridstiami dvoma pri enkódovaní žánru.

```
model = Sequential()
model.add(Dense(1024, input_dim=X_train.shape[1], activation=
    'relu'))
model.add(Dense(1024, activation='relu'))
model.add(Dense(512, activation='relu'))
model.add(Dense(y_train.shape[1], activation='softmax'))

model.compile(loss='categorical_crossentropy', optimizer=Adam
    (), metrics=['accuracy'])

history = model.fit(x=X_train, y=y_train, validation_data=(
    X_valid, y_valid), epochs=100, batch_size=32)
```

Náš model je husto prepojená neurónová sieť navrhnutá pre problém klasifikácie viacerých tried. Tu je popis modelu a odôvodnenie vykonaných volieb:

- Model pozostáva z troch husto spojených (tiež známych ako plne spojené) skrytých vrstiev s 1024, 1024 a 512 neurónmi. ReLU (Rectified Linear Unit) je zvolený pre svoju efektívnosť pri nelineárnych transformáciách v neurónových sieťach.
- Konečná vrstva má počet neurónov rovnajúcim sa y_train.shape, čo zodpovedá počtu tried v cieľovej premennej. activation='softmax': Softmax sa používa ako aktivačná funkcia vo výstupnej vrstve pri problémoch s klasifikáciou viacerých tried.
- loss='categorical_crossentropy': Táto funkcia straty sa používa, ak existujú dve alebo viac tried pri klasifikácii. Očakáva OneHotEncoding enkódovanie. Meria výkonnosť klasifikačného modelu, ktorého výstupom je hodnota pravdepodobnosti medzi 0 a 1.
- optimalizátor=Adam(): Adam je optimalizačný algoritmus, ktorý možno použiť namiesto klasického stochastického gradientu na iteratívnu aktualizáciu váh siete na základe tréningových údajov.
- 100 epoch: Veľký počet epoch bez implementácie predčasného zastavenia alebo iných techník regularizácie (ako je vypadnutie alebo pokles váhy) je zámerný na vyvolanie pretrénovania. K pretrénovaniu dochádza, keď sa model učí tréningové údaje príliš dobre, zachytáva šum a detaily, ktoré sa nezovšeobecňujú na nové údaje.

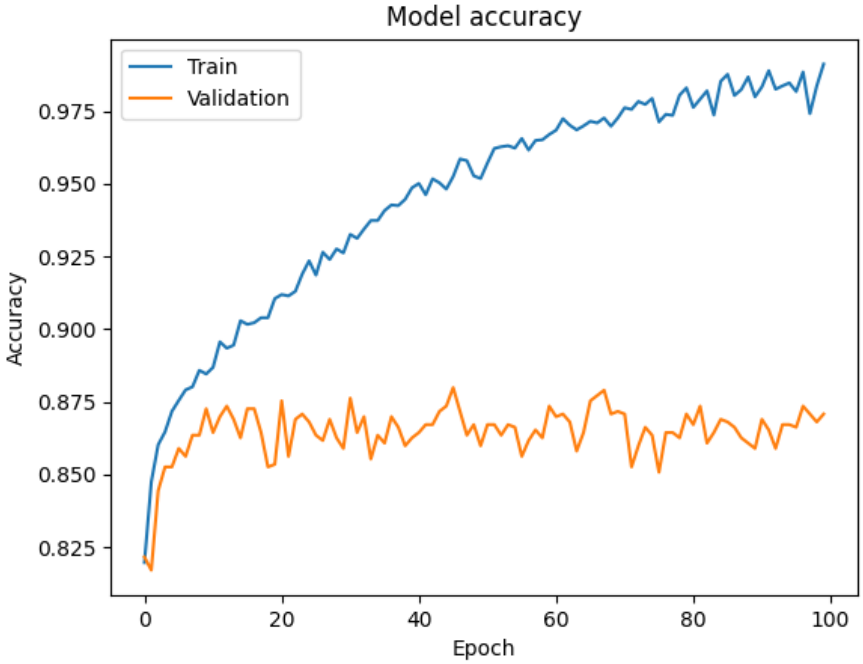

Figure 37: Model Accuracy Graph

- Presnosť tréningu (modrá čiara) sa neustále zvyšuje s rastúcim počtom epoch, čo naznačuje, že model sa učí z tréningových údajov. Presnosť validácie (oranžová čiara) sa tiež spočiatku zvyšuje, ale neskôr začína kolísať. Toto správanie naznačuje, že model sa nezovšeobecňuje tak ako by mal.
- Strata tréningu (modrá čiara) klesá, čo sa očakáva, keď sa model učí. Strata validácie (oranžová čiara) spočiatku klesá, ale potom sa začína zvyšovať, čo je klasický znak pretrénovania. Model funguje dobre s tréningovými údajmi, ale nedokáže udržať tento výkon na údajoch, ktoré predtým nevidel.
- Confusion matrix poskytuje rozdelenie predpovedí vs. správnych odpovedí. Diagonálne prvky predstavujú počet bodov, pre ktoré sa predpokladaná hodnota rovnala pravdivej. Pretrénovaný model bude mať často vysoký počet správnych predpovedí na tréningovej súprave, ale môže vykazovať viac nesprávnych klasifikácií validácie alebo testovacej sady.
- Toto meranie malo test accuracy: 0.8664

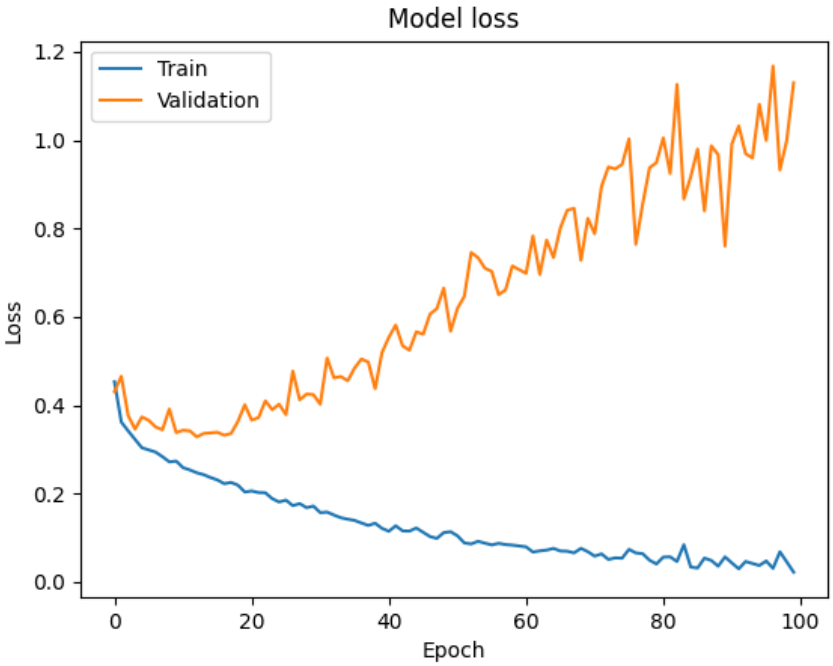

Figure 38: Model Loss Graph

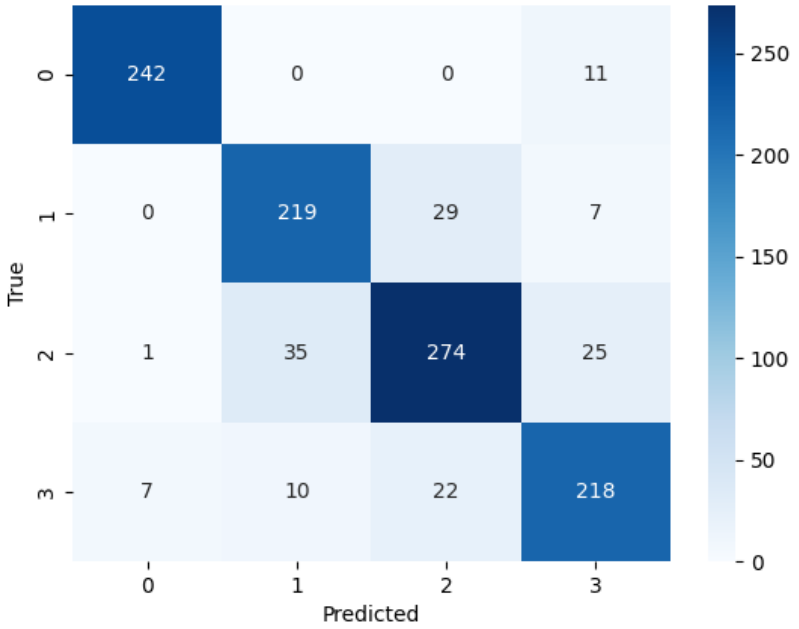

Figure 39: Confusion matrix pre testovaciu množinu

5.2 EarlyStopping

Ak chcete zmierniť nadmerné vybavenie, môžete zaviesť EarlyStopping(predčasné zastavenie) počas tréningu. Predčasné zastavenie je forma regularizácie, ktorá sa používa na zabránenie nadmernému vybaveniu zastavením tréningového procesu, ak sa výkon modelu na validačnej sade prestane zlepšovať na určitý počet epoch, známy ako trpezlivosť.

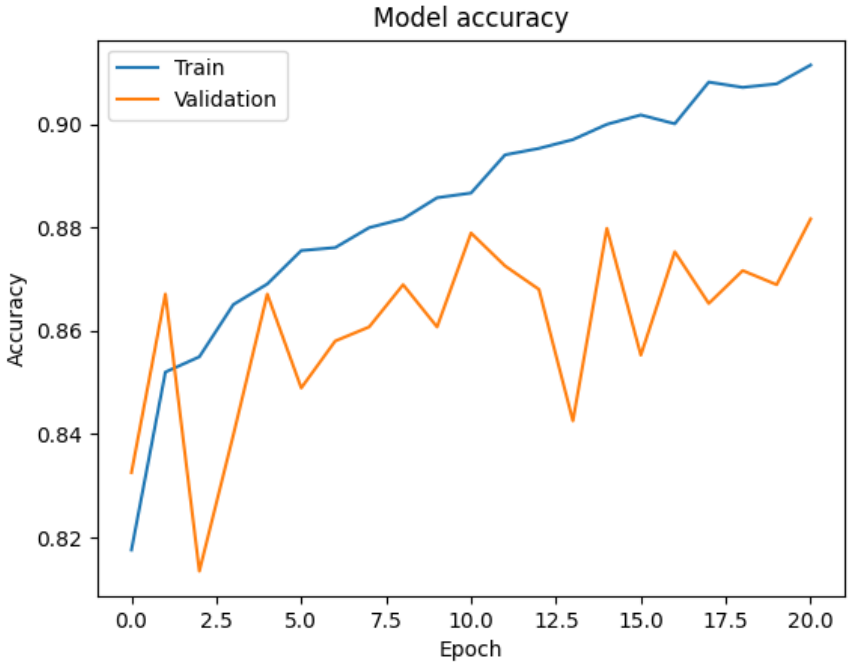

Figure 40: Model Accuracy Graph - EarlyStopping

- Presnosť tréningu sa časom zvyšuje, čo naznačuje, že model sa učí a zlepšuje svoj výkon na základe tréningových údajov. Presnosť validácie je trochu nestála a nevykazuje konzistentný nárast. Kolíše, ale neexistuje jasný trend zhoršovania v priebehu času, čo je dobrým znamením, že model nie je výrazne nadmerne fit.
- Strata tréningu sa neustále znižuje, čo sa očakáva počas tréningu, pretože model sa učí lepšie vyhovovať tréningovým údajom. Validačná strata po počiatočnom znížení vykazuje určitú volatilitu a mierne zvýšenie v neskorších epochách.
- Test accuracy: 0.8573, čo je zaujímavý výsledok. Taktiež aj v konfúznej matici vidíme, že pretrénovaný model mal stále o niečo vyššiu presnosť ako model s EarlyStoppingom. Keďže presnosť pre pretrénovaný model neklesala prudko, ale oscilovala, tento jav nie je až taký prekvapujúci. Nastavením inej trpezlivosti by sme dosiahli iný výsledok.

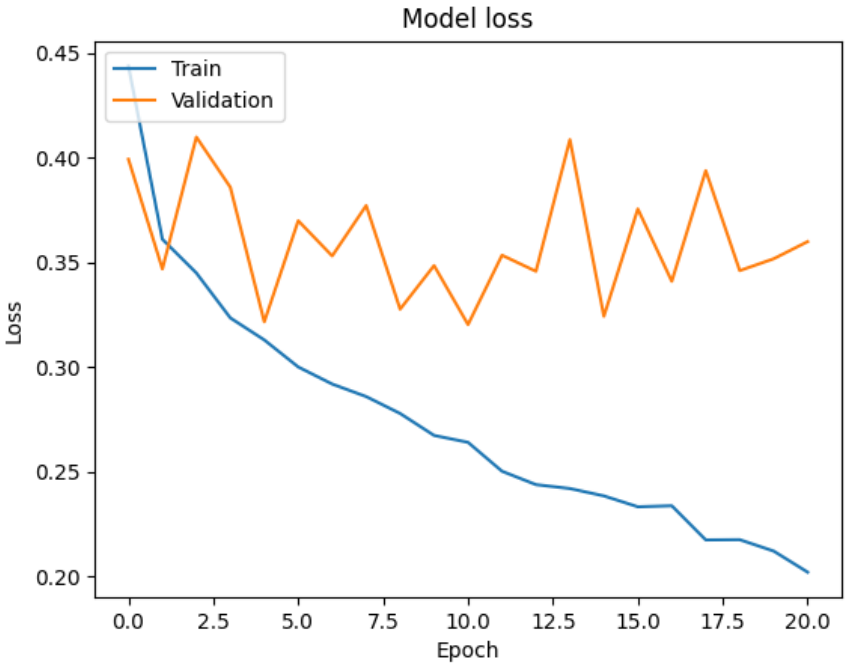

Figure 41: Model Loss Graph - EarlyStopping

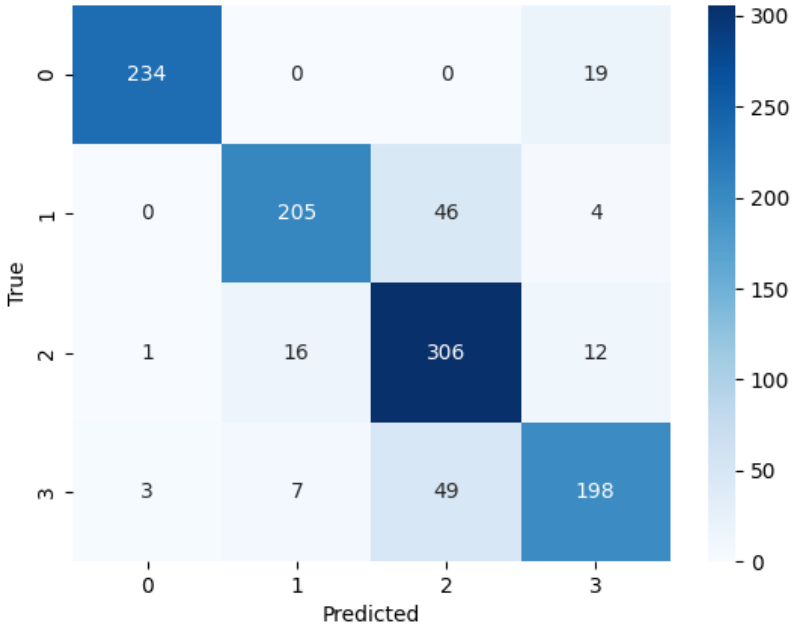

Figure 42: Confusion matrix pre testovaciu množinu - EarlyStopping

5.3 Testovanie

Index	Zmenený parameter	Úspešnosť - train	Úspešnosť -test
1	learning_rate=0.01	0.8936	0.8400
2	learning_rate=0.0001	0.9147	0.8500
3	learning_rate=0.005	0.8873	0.8555
4	Neurónov v 2. vrstve: 1024 => 2048	0.8965	0.8518
5	Neuróny v 3. vrstve: 512 => 256	0.9171	0.8500

Table 2: Tabuĺka s výsledkami testovania

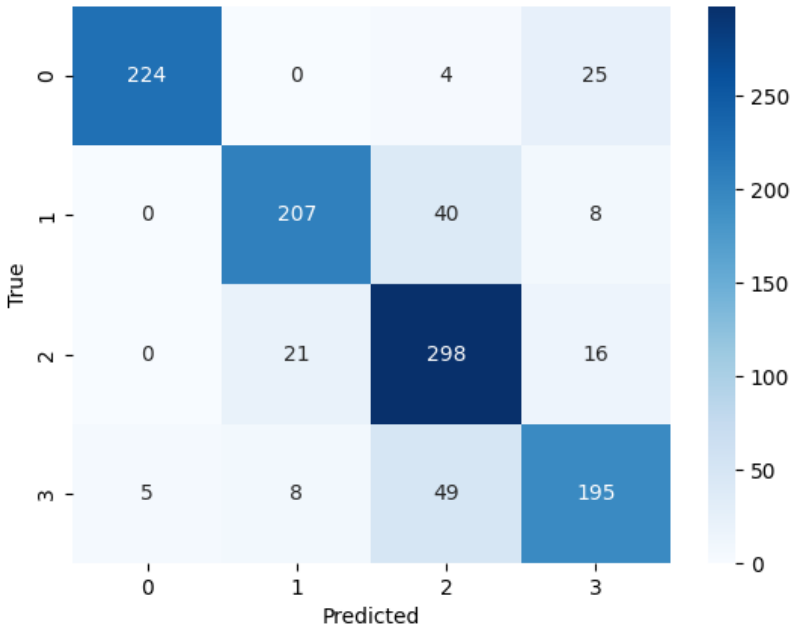

Figure 43: Model Accuracy Graph - Najhoršie meranie

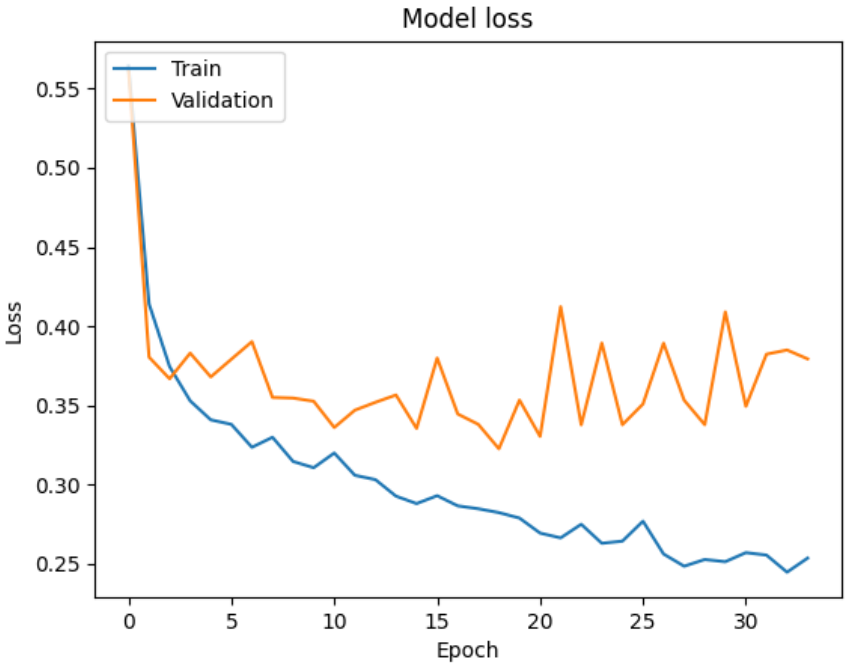

Figure 44: Model Loss Graph - Najhoršie meranie

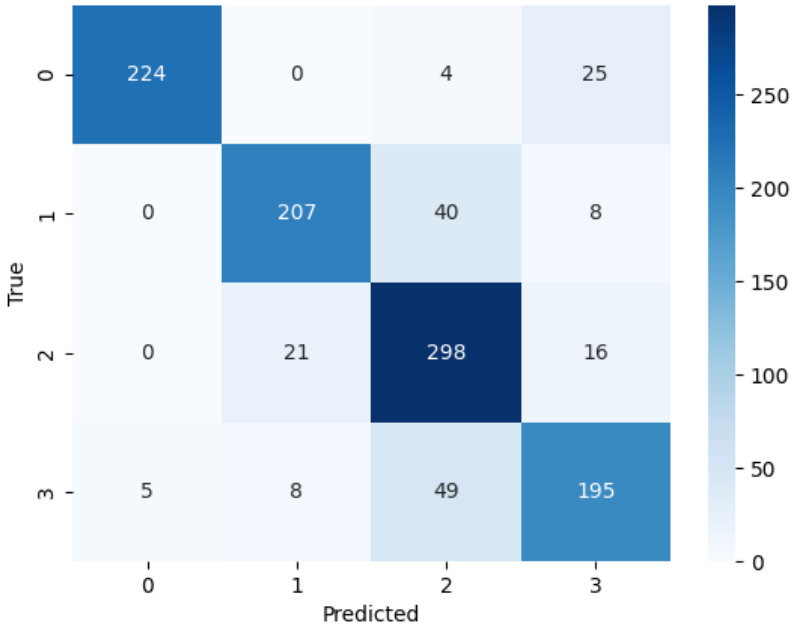

Figure 45: Confusion Matrix - Najhoršie meranie

6 References

- Materiály a kódy k predmetu SUNS na FEI STU.
- Informácie z internetového kurzu
- S metódou odstránenia outlierov IQR a vykreslením grafom so seaborn mi pomohol ChatGPT