1. Uvod i Opis Projekta

1. Uvod i Opis Projekta

1.1 Cilj projekta

Ovaj projekat predstavlja sveobuhvatnu analizu podataka o trending YouTube video snimcima korišćenjem tehnika klasterovanja. Glavni cilj je otkrivanje prirodnih grupa video snimaka na osnovu njihovih karakteristika kao što su angažovanje korisnika, temporalni obrasci i sadržajne osobine.

1.2 Dataset

- **Izvor:** Kaggle - Trending YouTube Video Statistics
- **Fajl:** DEvideos.csv (Nemački YouTube trending video snimci)
- **Veličina:** 29,627 video snimaka nakon uklanjanja duplikata
- **Broj atributa:** 102

1.3 Tehnologije

- Python 3.11
- Biblioteke: pandas, numpy, scikit-learn, matplotlib, seaborn, kagglehub

---

2. Metodologija

2.1 Preprocesiranje podataka

Pipeline za preprocesiranje uključuje:

- Uklanjanje duplikata (uklonjeno ~11,000 duplikata)
- Rukovanje nedostajućim vrednostima
- Konverzija datumskih kolona
- Uklanjanje irelevantnih kolona (video_id, thumbnail_link, itd.)

2.2 Karakteristike

Kreirano je preko 100 atributa iz originalnih podataka:

Temporalne karakteristike:

- Dan u nedelji, mesec, sat objavljivanja
- Trajanje trendovanja (trending_duration)
- Indikator vikenda

Metrike angažovanja:

- Lajkovi po pregledu (likes_per_view)
- Komentari po pregledu (comments_per_view)
- Stopa angažovanja (engagement_rate)
- Odnos lajkova i dislajkova (like_ratio)

Tekstualne karakteristike:

- Dužina naslova i broj reči
- Sentiment naslova
- Broj tagova
- TF-IDF reprezentacija tagova (50 atributa)

Kategorijske karakteristike:

- One-hot enkodiranje kategorija video snimaka

2.3 Algoritmi klasterovanja

Implementirano je 5 algoritama:

1. **KMeans** - Particiono klasterovanje sa elbow metodom za optimalni k
2. **Aglomerativno klasterovanje** - Hijerarhijsko klasterovanje sa Ward vezom
3. **DBSCAN** - Klasterovanje zasnovano na gustini
4. **Gaussian Mixture Models (GMM)** - Probabilističko klasterovanje
5. **Spektralno klasterovanje** - Klasterovanje zasnovano na grafovima

2.4 Redukcija dimenzionalnosti

Modeli su trenirani na tri varijante skupa podataka:

- **Full features** - Svih 102 atributa
- **PCA** - Redukcija sa zadržavanjem 95% varijanse
- **SelectKBest** - Top 50 atributa po F-score

---

3. Rezultati

3.1 Metrike evaluacije

Algoritam	Skup	Silhouette	Davies-Bouldin	Calinski-Harabasz
KMeans	full	0.0665	2.6497	1093.79
Aglomerativno	full	**0.2583**	1.8254	976.47
DBSCAN	full	0.3617	**0.7561**	365.16

Algoritam	Skup	Silhouette	Davies-Bouldin	Calinski-Harabasz
GMM	full	0.0101	5.8799	458.20
Spektralno	full	0.1658	2.5199	487.54

Napomena: Viši Silhouette i Calinski-Harabasz su bolji, niži Davies-Bouldin je bolji.

3.2 Najbolji model

- **Algoritam:** Aglomerativno klasterovanje
- **Skup karakteristika:** Full features (svi atributi)
- **Silhouette Score:** 0.2583
- **Davies-Bouldin Index:** 1.8254
- **Calinski-Harabasz Score:** 976.47

3.3 Identifikovani klasteri

Klaster 1: Mainstream (93.0% - 27,540 video snimaka)

Ovaj klaster predstavlja tipične trending video snimke sa prosečnim karakteristikama. Čini ogromnu većinu dataset-a i predstavlja "normalne" trending video snimke bez ekstremnih vrednosti.

Klaster 2: Niche Music Viral (0.3% - 79 video snimaka)

Ekstremno viralni video snimci sa izuzetno visokim metrikama:

- Pregledi: +13.05σ iznad proseka
- Ukupne interakcije: +13.04σ iznad proseka

Ovo su mega-viralni muzički video snimci koji su daleko nadmašili sve ostale po broju pregleda i interakcija.

Klaster 3: Film Specialized (6.4% - 1,883 video snimaka)

Video snimci specijalizovani za filmski sadržaj sa karakterističnim tagovima. Imaju specifičan profil tagova koji ih jasno razlikuje od ostalih kategorija.

Klaster 4: Niche People Specialized (0.4% - 125 video snimaka)

Mali ali distinktivan klaster iz kategorije "People & Blogs" (Kategorija 22):

- Specifični tagovi: +15.36σ iznad proseka
- Naslovi sa brojevima: +0.89σ iznad proseka

---

4. Vizualizacije

Projekat generiše sledeće vizualizacije:

1. **clusters_2d.png** - 2D PCA projekcija klastera
2. **clusters_3d.png** - 3D PCA projekcija klastera
3. **cluster_distribution.png** - Distribucija veličina klastera
4. **cluster_characteristics.png** - Heatmapa karakteristika klastera
5. **correlation_heatmap.png** - Korelaciona matrica atributa
6. **elbow_curve.png** - Elbow metoda za optimalni broj klastera
7. **metrics_comparison.png** - Poređenje metrika algoritama

---

5. Zaključci

5.1 Glavni nalazi

1. **Dominacija mainstream sadržaja:** 93% trending video snimaka ima slične, prosečne karakteristike, što ukazuje da većina trending sadržaja prati sličan obrazac.
2. **Postojanje viralnih outlier-a:** Mali broj video snimaka (0.3%) ima ekstremno visoke metrike - to su mega-viralni snimci koji se značajno razlikuju od ostatka.
3. **Specijalizacija po sadržaju:** Klasteri se jasno razlikuju po tipu sadržaja (muzika, film, blogovi), što potvrđuje da različite kategorije imaju različite obrasce angažovanja.
4. **Aglomerativno klasterovanje je najefikasnije:** Od 5 testiranih algoritama, aglomerativno klasterovanje sa Ward vezom daje najbolje rezultate za ovaj dataset.

5.2 Uticaj redukcije dimenzionalnosti

- **PCA:** Blago poboljšava KMeans i GMM, ali smanjuje performanse aglomerativnog klasterovanja
- **SelectKBest:** Značajno poboljšava Calinski-Harabasz score za većinu algoritama

5.3 Preporuke

1. Za analizu YouTube trending video snimaka preporučuje se korišćenje **aglomerativnog klasterovanja sa svim atributima**.
2. Za brže procesiranje velikih dataset-a, **PCA redukcija** je prihvatljiva alternativa sa minimalnim gubitkom kvaliteta.
3. **DBSCAN** je koristan za identifikaciju outlier-a (šuma), ali stvara previše klastera za praktičnu upotrebu.

---

6. Struktura projekta

```
app/
├── data/
│   ├── raw/              # Sirovi podaci
│   └── processed/        # Procesirani podaci
├── models/               # Sačuvani modeli
```

```
├── results/              # Rezultati evaluacije
├── visualizations/       # Generisane vizualizacije
└── src/
    ├── main.py                 # Glavni skript
    ├── data_preprocessing.py   # Preprocesiranje
    ├── feature_engineering.py  # Inženjering karakteristika
    ├── clustering.py           # Algoritmi klasterovanja
    ├── evaluation.py           # Metrike evaluacije
    ├── visualization.py        # Vizualizacije
    ├── cluster_naming.py       # Imenovanje klastera
    └── download_dataset.py     # Preuzimanje podataka
```

---

7. Pokretanje projekta

```
# Instalacija zavisnosti
pip install -r requirements.txt

# Pokretanje
cd app/src
python main.py
```

Dataset se automatski preuzima sa Kaggle-a prilikom prvog pokretanja.