**Report sull'attività di Data Science con i dataset Prime Video e Netflix**

Link video: <https://youtu.be/UzDhYmVrk-w>

**Descrizione dell'attività svolta**

Nell'ambito di questa attività di Data Science, ho lavorato sui dataset di Prime Video e Netflix al fine di analizzare e stimare la popolarità dei contenuti presenti su entrambe le piattaforme. L'obiettivo principale era utilizzare le informazioni disponibili nei dataset per estrarre insight interessanti e sviluppare modelli predittivi per stimare la popolarità dei contenuti.

**Dataset**

I dataset di Prime Video e Netflix sono composti da diverse informazioni relative ai contenuti disponibili sulle rispettive piattaforme di streaming. Ogni dataset contiene queste colonne:

id

title

type

release\_year

age\_certification

runtime

genres

production\_countries

imdb\_score

tmdb\_popularity

tmdb\_score

**Obiettivo dell'attività di Data Science**

L'obiettivo principale dell'attività è stato quello di sviluppare modelli predittivi per stimare la popolarità dei contenuti su Prime Video e Netflix. Utilizzando le caratteristiche disponibili nei dataset, come i punteggi IMDB e TMDB e la durata, ho cercato di creare modelli in grado di fare previsioni accurate sulla popolarità dei contenuti.

**Attività svolta**

Per raggiungere l'obiettivo prefissato, ho seguito i seguenti passaggi:

1. Esplorazione dei dati: Ho analizzato i dataset di Prime Video e Netflix per comprendere la struttura e il contenuto dei dati. Ho identificato le variabili rilevanti per il mio obiettivo e ho eseguito una pulizia dei dati per gestire eventuali valori mancanti o anomali.

A screen shot of a computer code

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screen shot of a computer code

Description automatically generated with medium confidence

A picture containing text, screenshot, software, font

Description automatically generated

A picture containing text, screenshot, diagram, line

Description automatically generated

2. Preparazione dei dati: Ho effettuato una fase di pre-elaborazione dei dati, che ha incluso la quantizzazione della colonna tmdb\_popularity in 3 macro gruppi: low, medium, high.

A picture containing screenshot, text, font, line

Description automatically generated

3. Selezione del modello: Ho scelto di utilizzare un modello di regressione logistica multiclasse per creare i modelli predittivi per stimare la popolarità dei contenuti.

A picture containing text, screenshot, font

Description automatically generated

4. Addestramento del modello: Ho suddiviso i dataset in set di addestramento e set di test. Ho utilizzato come set di addestramento quello di Prime Video e come set di test quello di Netflix.

A screen shot of a computer

Description automatically generated with low confidence

A picture containing text, screenshot, diagram, colorfulness

Description automatically generated

Accuratezza: 0.45

A picture containing text, screenshot, colorfulness, diagram

Description automatically generatedA picture containing text, screenshot

Description automatically generated

Accuratezza: 0.39

5. Valutazione dei risultati: Ho valutato le prestazioni dei modelli utilizzando l'accuratezza e ho dedotto che il modello non si comporta egregiamente. Mantiene le promesse, ovvero l’accuratezza è più o meno la stessa per i 2 dataset ma il lato negativo è che è troppo bassa. Ho considerato le limitazioni e le aree di miglioramento:

LIMITAZIONI:

1. Il descrittore non è composto alla perfezione: Sebbene le colonne di runtime, imdb\_score e tmdb\_score siano informative, da sole potrebbero non essere sufficienti per predire accuratamente la popolarità di un film o show.
2. Dimensione limitata del dataset: Il dataset utilizzato potrebbe non essere estremamente ampio, il che potrebbe influire sulla capacità dei modelli di fornire stime accurate. L'aggiunta di più dati potrebbe migliorare la capacità predittiva dei modelli.

AREE DI MIGLIORAMENTO:

1. Aggiungere il genere come colonna: L'aggiunta del genere come attributo nel descrittore potrebbe contribuire a migliorare la capacità predittiva del modello. Il genere può essere un fattore importante nell'influenzare la popolarità di un film o show, poiché i gusti degli spettatori possono variare in base al genere preferito.
2. Includere la certificazione per l'età come colonna: La certificazione per l'età può essere un'altra caratteristica significativa per stimare la popolarità dei contenuti. La preferenza degli spettatori può essere influenzata dalla fascia di età a cui il contenuto è destinato; quindi, includere la certificazione come attributo potrebbe aiutare a migliorare le previsioni.
3. Esplorare ulteriori attributi: È consigliabile esaminare altri attributi disponibili nei dataset o cercare di acquisire ulteriori dati rilevanti. Ad esempio, considerare le recensioni degli utenti, il numero di visualizzazioni, la presenza di attori famosi o la copertura mediatica potrebbe fornire informazioni aggiuntive per migliorare la stima della popolarità.
4. Utilizzare modelli più complessi: Oltre alla regressione logistica, potrebbe essere utile esplorare altri modelli più complessi, come alberi di decisione, support vector machines o modelli di apprendimento profondo. Questi modelli potrebbero catturare relazioni più complesse tra gli attributi e la popolarità dei contenuti.
5. Raccolta di dati aggiuntivi: Se possibile, considerare la raccolta di ulteriori dati da diverse fonti, come informazioni sul marketing o sulla distribuzione dei contenuti. Questi dati possono fornire ulteriori dettagli sul contesto di lancio e promozione dei film o show, che potrebbero influenzare la popolarità.

**Risultati ottenuti**

I modelli di regressione logistica hanno fornito risultati promettenti nella stima della popolarità dei contenuti su Prime Video e Netflix. L'accuratezza dei modelli è stata del 39% nel set di test. Tuttavia, è importante notare che la popolarità dei contenuti può essere influenzata da diversi fattori, come la promozione, il genere, la certificazione dell’età e le tendenze di visualizzazione, che potrebbero rendere difficile una previsione accurata.

**Conclusioni**

In conclusione, l'attività di Data Science utilizzando i dataset di Prime Video e Netflix è stata interessante e ha fornito risultati non promettenti poiché la popolarità è alquanto difficile da stimare con i dati e i modelli scelti. Tuttavia, è importante considerare che i modelli sviluppati potrebbero beneficiare di ulteriori miglioramenti e affinamenti. L'analisi dei dati relativi ai contenuti su Prime Video e Netflix può essere utile per ottenere insight sulle preferenze degli utenti e identificare i fattori che influenzano la popolarità dei contenuti.

**Risorse utilizzate**

Durante lo svolgimento di questa attività di Data Science, ho utilizzato diverse risorse, tra cui:

- Dataset di Prime Video e Netflix: <https://github.com/mattiacentonze/data_science_one-o-one>

- Librerie utilizzate:

import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

import scipy.stats as stats

from sklearn.metrics import accuracy\_score

from sklearn.model\_selection import train\_test\_split

from sklearn.decomposition import PCA

from sklearn.preprocessing import StandardScaler

from sklearn.linear\_model import LogisticRegression

from mlxtend.plotting import plot\_decision\_regions

- Documentazione e tutorial online per il supporto tecnico e l'apprendimento delle tecniche di machine learning utilizzate:

- Materiale didattico su Aulaweb del corso Data Science one-o-one

- <https://www.andreaminini.com/python/sklearn/regressione-logistica-con-python>

- Introduzione alla Data Science – Corso terzo anno Informatica Triennale