Immagine che contiene testo, screenshot, monitor

Descrizione generata automaticamente

**FILMTV MOVIES**  
Documentazione Progetto

Corso di Ingegneria della Conoscenza

A.A. 2022-2023

Gruppo di lavoro

* Nicola Dipinto, 729920, [n.dipinto8@studenti.uniba.it](mailto:n.dipinto8@studenti.uniba.it)
* Andrea Biasco, 744605, [a.biasco3@studenti.uniba.it](mailto:a.biasco3@studenti.uniba.it)

URL repo associato: <https://github.com/nicola1997/icon22-23.git>

Indice

[INTRODUZIONE 3](#_Toc127225050)

[Sommario 3](#_Toc127225051)

[Elenco argomenti di interesse 3](#_Toc127225052)

[Strumenti adottati 3](#_Toc127225053)

[PREPROCESSING DEI DATI 4](#_Toc127225054)

[Sommario 4](#_Toc127225055)

[Decisioni di Progetto 4](#_Toc127225056)

[Valutazione 5](#_Toc127225057)

[CLUSTERING 6](#_Toc127225058)

[Sommario 6](#_Toc127225059)

[Decisioni di Progetto 6](#_Toc127225060)

[Valutazione 6](#_Toc127225061)

[KNOWLEDGE BASE 7](#_Toc127225062)

[Sommario 7](#_Toc127225063)

[Decisioni di Progetto 7](#_Toc127225064)

[Valutazione 7](#_Toc127225065)

[BAYESIAN NETWORK 7](#_Toc127225066)

[Sommario 7](#_Toc127225067)

[Decisioni di Progetto 7](#_Toc127225068)

[Valutazione 7](#_Toc127225069)

[Conclusioni 9](#_Toc127225070)

[Riferimenti Bibliografici 9](#_Toc127225071)

INTRODUZIONE

Gli obiettivi del nostro progetto sono molteplici e ciascuno relativo all’utilizzo che si è fatto dei vari metodi di apprendimento. In questo progetto abbiamo esplorato l’utilizzo dell’apprendimento automatico per analizzare un dataset “[FILMTV MOVIES](https://www.kaggle.com/datasets/stefanoleone992/filmtv-movies-dataset?select=filmtv_movies+-+ITA.csv)”.

Sommario

Attraverso il **Preprocessing** dei dati, abbiamo preparato le informazioni per l’elaborazione e l’analisi successiva. In seguito, abbiamo utilizzato tecniche di **Clustering** per raggruppare i film in base a determinare caratteristiche. Inoltre, abbiamo creato una **Knowledge Base** per rappresentare le informazioni sul dataset in modo coerente e accessibile. Infine, abbiamo applicato la **Bayesian Network** per effettuare previsioni sulle caratteristiche dei film e per comprendere le relazioni causali tra le diverse variabili.

In questo modo siamo stati in grado di ottenere una comprensione più approfondita del dataset e di utilizzare l’apprendimento automatico per sviluppare nuove conoscenze e previsioni sul mondo dei film.

Elenco argomenti di interesse

* **PREPROCESSING DEI DATI:** PCA, MinMaxScaler
* **CLUSTERING:** utilizziamo l’algoritmo K-Means per la creazione di cluster.
* **KNOWLEDGE BASE**: utilizziamo Prolog per il ragionamento su una base di conoscenza partendo dai dati contenuti nel dataset, permettendo di inferire nuove informazioni.
* **BAYESIAN NETWORK:** usiamo la tecnica Variable Elimination, MaxmumlikelihoodEstimator.

Strumenti adottati

Il linguaggio utilizzato per sviluppare il progetto è **Python**, data la grande potenzialità nel manipolare dati con le numerose librerie come **Pandas** per la gestione del dataset, **PySwip** per poter usare i comandi dell’applicativo **SWI-Prolog**, **Sklearn** per la parte relativa al clustering e **Pgmpy** per l’inferenza probabilistica effettuata con la Bayesian Network. **Matplotlib** per visualizzare i dati dei film, **Numpy** per lavorare con array. **Sys** per la gestione degli argomenti da riga di comando.

# PREPROCESSING DEI DATI

## Sommario

Il **preprocessing** del dataset dei film è una fase fondamentale nell’elaborazione delle informazioni. Serve per pulire, trasformare e preparare i dati in modo che siano adatti all’analisi e all’utilizzo successivo. Il preprocessing è stato effettuato prima di applicare il Clustering, Knowledge Base e la Bayesian Network perché abbiamo bisogno di dati precisi e fornire dati affidabili. Grazie al preprocessing abbiamo rimosso eventuali valori mancanti o duplicati, normalizzare i valori e discretizzare.

## Decisioni di Progetto

1. Inizialmente abbiamo dato una **pulizia generale** al dataset filmtv.csv: abbiamo deciso di eliminare le colonne inutili ai nostri scopi finali come 'voto\_critica','voto\_pubblico','voti\_totali','descrizione','note','titolo\_italiano' e ci siamo accertati che nel dataset non ci fossero film duplicati o film con informazioni essenziali mancanti come ad esempio l’anno di uscita o la durata. In seguito, abbiamo fatto in modo di avere nelle colonne “attori”, “registi”, “paese” solo un valore, questo perché complicava le fasi successive del progetto.
2. Per quanto riguarda il **preprocessing per il cluster**, ci siamo limitati a considerare le principali feature per considerare la similarità dei film come genere, anno, durata, paese, humor, ritmo, impegno, tensione, erotismo. Per fare ciò è stato necessario applicare la Principal Component Analysis (**PCA**), per ridurre il numero di features, e uno dei metodi più utilizzati per scalare i valori dei features, ovvero **MinMaxScaler**: trasforma i dati nell’intervallo tra 0 e 1.
3. Dopo aver effettuato il clustering, si fa a definire il dataset che verrà utilizzato per **la Knowledge Base**, ottenuto semplicemente aggiungendo all’interno del dataset ottenuto dall’operazione di cleaning, una colonna con il relativo cluster per ciascun film.
4. Per quanto riguarda il **preprocessing per la** **Bayesian Network**, è stato necessario eliminare le colonne non funzionali all’inferenza probabilistica. La procedura importante è stata quella di discretizzazione di varie colonne, trasformate in valori testuali per rendere più comprensibile e naturale il linguaggio con cui l’utente andrà ad interagire col la Bayesian Network. I valori sono stati eliminati col l’utilizzo del metodo di inferenza “**Variable Elimination**” il quale richiamato sui valori non ancora discretizzati, va ad indicare gli intervalli per suddividere il dataset in maniera omogenea.

Per lo stesso motivo abbiamo cancellato gli spazi tra le stringhe e le abbiamo rese tutte minuscole

## Valutazione

Paragrafi che richiamino (non spieghino, se standard) le metriche adottate   
+ tabelle sui risultati e loro discussione

# CLUSTERING

## Sommario

Un paragrafo che chiarisca la rappresentazione della conoscenza scelta per KB (modelli di ragionamento / apprendimento), dati, BK, ...

## Decisioni di Progetto

Configurazione dei componenti (e.g. API/librerie) utilizzati, ad es. parametri, soglie, ecc.   
e di eventuali metodi specifici utilizzati a tale scopo

## Valutazione

Paragrafi che richiamino (non spieghino, se standard) le metriche adottate   
+ tabelle sui risultati e loro discussione

# KNOWLEDGE BASE

## Sommario

Un paragrafo che chiarisca la rappresentazione della conoscenza scelta per KB (modelli di ragionamento / apprendimento), dati, BK, ...

## Decisioni di Progetto

Configurazione dei componenti (e.g. API/librerie) utilizzati, ad es. parametri, soglie, ecc.   
e di eventuali metodi specifici utilizzati a tale scopo

## Valutazione

Paragrafi che richiamino (non spieghino, se standard) le metriche adottate   
+ tabelle sui risultati e loro discussione

# BAYESIAN NETWORK

## Sommario

Una Bayesian Network è un modello che rappresenta in forma grafica una distribuzione di probabilità congiunta su più variabili, alcune delle quali sono dipendenti l’una dalle altre. Tale modello è rappresentato da un grafo orientato aciclico il quale indica il modo in cui le feature sono tra loro dipendenti.

## Decisioni di Progetto

Per costruire la rete Bayesiana abbiamo deciso quali features usare per i nodi del grafo. Successivamente si individuano i nodi genitori e i nodi radice. A partire dai nodi-radice si può sviluppare la rete Bayesiana seguendo le relazioni padre figlio fino agli ultimi nodi.

Abbiamo deciso di utilizzare la Bayesian Network per analizzare i dati e scoprire relazioni tra variabili come: voto, anno, paese, durata, genere. Inoltre, permette all’utente di eseguire interrogazioni basate sull’inferenza probabilistica.

# l'algoritmo MaximumLikelihoodEstimator viene utilizzato per addestrare la rete bayesiana. Questo algoritmo utilizza l'equazione del massimo verosimiglianza per calcolare le probabilità condizionali dei nodi nella rete.

## Valutazione

Paragrafi che richiamino (non spieghino, se standard) le metriche adottate   
+ tabelle sui risultati e loro discussione

# Conclusioni

Un paragrafo che riassuma le valutazioni e delinei possibili sviluppi, ad. es. problematiche non affrontate per questioni di tempo (per eventuali estensioni da parte di altri gruppi).

# Riferimenti Bibliografici

[1] ...

[2] ...

[3] ...