

Progetto di Data Analytics  
A.A. 2022-2023

Daniele Perrella 1038893  
Simone Boldrini 1038792

13 marzo 2023

# Indice

<b>1</b>	<b>Introduzione</b>	<b>2</b>
<b>2</b>	<b>Metodologia</b>	<b>3</b>
2.1	Data Acquisition . . . . .	3
2.2	Dataset . . . . .	3
2.3	Data Visualization . . . . .	5
2.4	Tabular Data . . . . .	5
<b>3</b>	<b>Implementazione</b>	<b>6</b>
3.1	Data Preprocessing . . . . .	6
3.2	Modeling . . . . .	6
<b>4</b>	<b>Risultati</b>	<b>7</b>
4.1	Performance Evaluation . . . . .	7
4.2	Performance Evaluation on Tabular Data . . . . .	7

# Capitolo 1

## Introduzione

L'obiettivo di questo report è presentare il progetto del corso di Data Analytics finalizzato alla predizione del voto medio di un film, date le sue caratteristiche, utilizzando un dataset proveniente da MovieLens, un recommendation system per contenuti video. Il dataset contiene rating e tag per oltre 60.000 film, raccolti da più di 150.000 utenti negli anni 1995-2019. Ogni file del dataset dispone di un genoma che identifica una caratteristica del film e la sua rilevanza.

Per raggiungere questo obiettivo, abbiamo utilizzato tecniche di Machine Learning supervisionate tradizionali quali Linear Regression, SVM, NB in seguito verranno illustrate nel dettaglio; tecniche di ML basate su Reti Neurali ed infine modelli deep per Tabular Data. In particolare, il report descrive il processo di acquisizione e preparazione dei dati, l'analisi esplorativa del dataset, la selezione delle feature rilevanti e la costruzione dei modelli predittivi.

Infine, il report conclude con una valutazione critica dei modelli creati, evidenziando i loro punti di forza e di debolezza, e suggerisce possibili sviluppi futuri per migliorare ulteriormente la predizione del voto medio dei film.

# Capitolo 2

## Metodologia

### 2.1 Data Acquisition

In uno studio di Data Analysis, lo step di Data Acquisition è il primo della pipeline da seguire. In questa fase abbiamo raccolto i dati necessari dal dataset [?, MovieLens]. I file sono stati raccolti in una cartella denominata `ml-25m` contenente un README.

### 2.2 Dataset

MovieLens è un Recommendation System per contenuti multimediali, quali film, serie tv, documentari ecc. MovieLens mette a disposizione degli sviluppatori un dataset open-source, generato dal database TMDB (The Movie Database). Questo dataset contiene recensioni con voti e tag di oltre 60.000 film, raccolte da oltre 150.000 utenti durante il periodo che va dal 1995 fino al 2019.

Il dataset è composto dai seguenti file:

- *genome-scores.csv*: contiene il *relevance score* di ogni tag per tutti i film (ovvero, quanto un tag è importante per il dato film).
- *genome-tags.csv*: contiene tutti i tag presenti all'interno del dataset.
- *links.csv*: contiene gli ID di ogni film per i due database TMDB e IMDB.
- *movies.csv*: contiene i titoli dei film, con i rispettivi generi.
- *ratings.csv*: contiene più di 25.000.000 votazioni provenienti dalle recensioni degli utenti.

- *tags.csv*: contiene i tag che sono stati assegnati ai film dagli utenti nelle rispettivi recensioni.

Per il nostro caso di studio, non abbiamo fatto uso di tutti i file disponibili, ma esclusivamente quelli che abbiamo considerato adatti per lo scopo. Di seguito viene mostrata la struttura dei tali.

Il dataset ***ratings*** è stato selezionato per poter far uso dei voti assegnati ai film dagli utenti, per poter quindi ottenere il voto medio di tali film. La struttura è la seguente

userId	movieId	rating	timestamp
1	296	5.0	1147880044
1	306	3.5	1147868817
1	307	5.0	1147868828
1	665	5.0	1147878820
1	899	3.5	1147868510
...	...	...	...

Tabella 2.1: ratings.csv

Il dataset ***genome-scores*** è stato selezionato per poter identificare le relazioni che ci sono tra un dato voto ed i valori dei tag assegnati

movieId	tagId	relevance
1	1	0.028749999999999998
1	2	0.023749999999999993
1	3	0.0625
1	4	0.075749999999999998
1	5	0.14075
...	...	...

Tabella 2.2: genome-scores.csv

Infine, si è fatto uso del dataset ***movies*** per ottenere le informazioni inerenti al genere di ogni film.

movieId	title	genres
1	Toy Story (1995)	Adventur Animation ...
2	Jumanji (1995)	Adventure Children ...
3	Grumpier Old Men (1995)	Comedy Romance
4	Waiting to Exhale (1995)	Comedy Drama ...
...	...	...

Tabella 2.3: movies.csv

## 2.3 Data Visualization

Nella fase di *data visualization* andremo a visualizzare alcune proprietà e peculiarità del dataset.

Distribuzione dei ratings

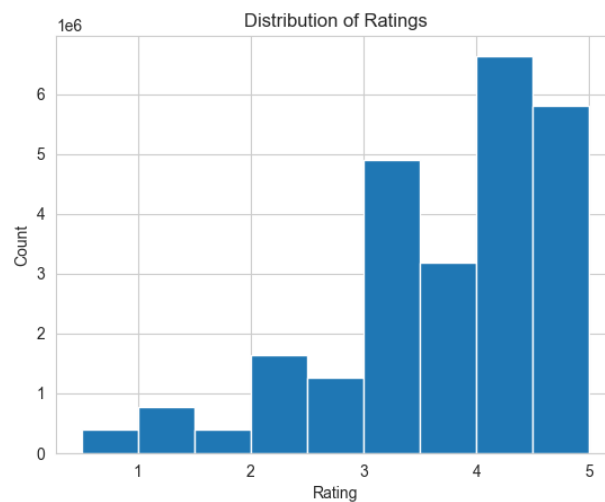


Figura 2.1: Distribution of Ratings.

## 2.4 Tabular Data

# Capitolo 3

## Implementazione

### 3.1 Data Preprocessing

La fase di preprocessing dei dati consiste nel preparare i dati per l'addestramento del modello. Una volta acquisiti i dati e averne studiato le peculiarità e le caratteristiche, si procede con la pulizia dei dati.

Abbiamo usato il metodo *pivot\_table* di Pandas per creare una tabella pivot che contiene i genome-score degli utenti per ogni film che lo possiede (andando così a ridurre la cardinalità da 60k a 13.816). In seguito a ciò verrà effettuata una merge con un i generi dei film applicando una One-Hot Encoding, attraverso la funzione *get\_dummies()*. Ci siamo inoltre andati a focalizzare sui voti da parte degli utenti, raggruppandoli per film ci abbiamo calcolato la media dei voti.

### 3.2 Modeling

#### Tabular Data

# Capitolo 4

## Risultati

### 4.1 Performance Evaluation

### 4.2 Performance Evaluation on Tabular Data