Progetto di Data Analytics
A.A. 2022-2023

Daniele Perrella 1038893
Simone Boldrini 1038792

13 marzo 2023

Indice

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

Il processo di Data Acquisition é la prima fase del nostro studio. In questa fase abbiamo raccolto i dati necessari dal dataset [**?**, MovieLens]. Attraverso questo sito abbiamo raccolto 6 file in formato CSV, che contengono informazioni su film, utenti e recensioni. I file sono stati raccolti in una cartella denominata `ml-25m` contente un README.

2.2 Data Visualization

2.3 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 acqusiti i dati e averne studiato le peculiaritá 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 cosi a ridurre la cardialita' da 60k a 13.816). In seguito a cio' verra' 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