Table of Contents

[Table of Figures 2](#_Toc198575145)

[Introduction 3](#_Toc198575146)

[ML – Part I – Clustering 4](#_Toc198575147)

[Feature Selection and Justification 4](#_Toc198575148)

[Clustering Algorithms 4](#_Toc198575149)

[Performance Evaluation 5](#_Toc198575150)

[Clustering Visualisation 5](#_Toc198575151)

[Interpretation of Results 8](#_Toc198575152)

[ML – Part II – Filtering Collaborative 8](#_Toc198575153)

[Feature Selection and Justification 8](#_Toc198575154)

[Applied Collaborative Filtering Algorithms 8](#_Toc198575155)

[Performance Evaluation and Best Approach 9](#_Toc198575156)

[Could Content-Based Filtering Improve Results? 9](#_Toc198575157)

[Interpretation of Results 10](#_Toc198575158)

[Data Visualisation 10](#_Toc198575159)

[Decision of Charts 10](#_Toc198575160)

[Genre Popularity vs. Average Rating Chart 10](#_Toc198575161)

[Top 5 Movies by Genre Chart 11](#_Toc198575162)

[Genre Trends Over Time Chart 11](#_Toc198575163)

[Top 10 Similar Movies by Genre Chart 12](#_Toc198575164)

[User Behaviour Analysis Radar Chart 12](#_Toc198575165)

[Dashboard Design 13](#_Toc198575166)

[Interactivity and user control 13](#_Toc198575167)

[Attractive but not overloaded visuals 13](#_Toc198575168)

[Clean and modern design 13](#_Toc198575169)

[Fast loading times 13](#_Toc198575170)

[Neutral language and inclusive visualisation 14](#_Toc198575171)

[Justification of the data preparation process for visualisations 14](#_Toc198575172)

[Conclusion 16](#_Toc198575173)

[References 17](#_Toc198575174)

[Streamlit Link 17](#_Toc198575175)

[Github Link 17](#_Toc198575176)

# 

# Table of Figures

[Figure 1 Elbow Method 5](#_Toc198575188)

[Figure 2 Dendrogram Plot 6](#_Toc198575189)

[Figure 3 K-Means Clusters (Genres Only) 7](#_Toc198575190)

[Figure 4 K-Means Clusters (Genres + Year) 7](#_Toc198575191)

[Figure 5 K-Medoids Clusters (Genres Only) 8](#_Toc198575192)

[Figure 6 Hierarchical Clustering (Genres Only) 8](#_Toc198575193)

# Introduction

This project explores two key machine learning approaches applied to the movie domain: unsupervised clustering to group movies according to genre similarities, and recommendation systems based on collaborative filtering and content.

Using a dataset of films, ratings and genres, pre-processing, coding and normalisation techniques were applied to prepare the data. Algorithms such as K-Means, K-Medoids and hierarchical clustering were implemented, as well as User-User, Item-Item and content-based recommendation models.

As a complement, an interactive dashboard was designed for a young audience (18-35 years old), with clear, dynamic and customisable visualisations. The overall goal was to improve the user experience by facilitating the exploration, discovery and recommendation of movies based on real data and scalable analytical approaches.

# ML – Part I – Clustering

## Feature Selection and Justification

The original dataset includes three columns: movieId, title and genres. MovieId was removed as it did not contain relevant information for the analysis. From title, the year of release was extracted using regular expressions, and a new title\_clean column containing the title without the year was created for easier reading.

The main feature used was genres, which includes several genres per film separated by the symbol ‘|’. This information was transformed by binary multi-label encoding using MultiLabelBinarizer, generating a column for each genre with values 1 or 0.

Subsequently, the year of release was incorporated as an additional variable in a second approach to assess whether the temporal dimension improves the quality of the clustering. All variables were normalised with StandardScaler to ensure that they have the same weight in the clustering algorithms.

## Clustering Algorithms

The following clustering algorithms were applied:

K-Means: This was used with genres features only and with genres plus year. The optimal number of clusters was estimated with the Elbow Method.

A graph with a line

AI-generated content may be incorrect.A graph with blue dots

AI-generated content may be incorrect.

Figure 1 Elbow Method

K-Medoids: Similar to K-Means but selects real points in the dataset as centroids (medoids), which makes it more robust to outliers.

Hierarchical Clustering (Agglomerative): A bottom-up technique was applied using the ward linkage criterion. A dendrogram was generated to visualise the hierarchy between the data and four cluster have been caught from the dendrogram.

A diagram of a city

AI-generated content may be incorrect.

Figure 2 Dendrogram Plot

To facilitate interpretation, Principal Component Analysis (PCA) was applied, and dimensionality was reduced to two components to generate visual plots of the clusters.

## Performance Evaluation

The Silhouette score was used as an evaluation metric. This metric measures how well a point fits within its own cluster compared to others. The results obtained were as follows:

|  |  |  |
| --- | --- | --- |
| Clustering Algorithm | Features | Silhouette Score |
| K-Means | Only Genres | 0.232 |
| K-Means | Genres + Year | 0.148 |
| K-Medoids | Only Genres | 0.203 |
| Hierarchical Clustering | Only Genres | 0.128 |

As can be observed, K-Means using only genres performed best (0.232), showing higher cohesiveness and separation among the clusters produced. The introduction of the year made no difference and decreased the quality of the clusters. K-Medoids was successful, although less so than K-Means, whereas hierarchical clustering was the least efficient.

## Clustering Visualisation

A diagram of a cluster of colorful dots

AI-generated content may be incorrect.

Figure 3 K-Means Clusters (Genres Only)

A screen shot of a computer screen

AI-generated content may be incorrect.

Figure 4 K-Means Clusters (Genres + Year)

A diagram of a cluster of dots

AI-generated content may be incorrect.

Figure 5 K-Medoids Clusters (Genres Only)

A diagram of a cluster of colored dots

AI-generated content may be incorrect.

Figure 6 Hierarchical Clustering (Genres Only)

## Interpretation of Results

The best performance was obtained by K-Means using only genders. However, a Silhouette Score of 0.232 is relatively low. This indicates that, although groups are identified, there is considerable overlap between clusters. One possible cause is the multi-label nature of the genre field, as many movies belong simultaneously to several genres (Comedy, Romance, and Drama’). This complicates the creation of clear boundaries between groups, as movies act as ‘bridges’ between clusters.

The inclusion of year of release did not add significant value; on the contrary, it dispersed the groupings, possibly because films from different eras share similar genres. K-Medoids, although more robust to outliers, did not outperform K-Means. Hierarchical clustering, while useful for visualising hierarchical relationships, did not achieve well-defined clusters.

In summary, it is recommended to use K-Means with genders as the only feature to perform clustering on this type of data. This approach offers a good balance between computational efficiency, visual interpretability and quality of results, being useful both for classification and for the development of content recommender systems.

# ML – Part II – Filtering Collaborative

## Feature Selection and Justification

For the developed recommender systems, the Movie and Ratings datasets, which include the columns userId, movieId, rating and genres, were used and merged. These features were selected for their relevance in both collaborative and content filtering methods.

In the collaborative filtering models, userId, movieId and rating were used, as they form the basis of the matrix of user-item interactions. Numerical ratings allow the calculation of similarities between users or between movies.

For the content-based model, the genres column was used. This structured semantic information allows to calculate similarities between films independently of user interactions.

In addition, the genres field was processed into lists (genres\_list) and then encoded with one-hot encoding using MultiLabelBinarizer, which facilitated the calculation of cosine similarity between movies according to their content.

## Applied Collaborative Filtering Algorithms

Three models were implemented:

a. Collaborative Item-Item Filtering (based on ratings)

A user-item matrix was built with the ratings and cosine similarity between movie vectors was calculated. From a given movie, the system recommends other movies with similar rating patterns by users. For example, for Toy Story (1995), films such as Star Wars (1977) and Jurassic Park (1993) were recommended, indicating a shared audience base.

b. User-User Collaborative Filtering (based on ratings)

The cosine similarity between users was calculated based on their ratings. For a target user, the most similar users are identified, and movies are recommended that they have rated positively but that the user has not yet seen. For example, for user 12882, films such as Blade Runner and Crouching Tiger, Hidden Dragon were recommended based on the likes of similar users.

c. Content-Based Filtering (based on genres)

This model calculates the similarity between movies based on their genres. Using the binarised genre matrix, films with similar thematic structure are identified. For Toy Story (1995), the system recommended films such as Antz, Toy Story 2 and Shrek, all with common genres such as Animation, Children's and Comedy.

## Performance Evaluation and Best Approach

A qualitative evaluation was conducted considering the relevance, diversity and personalisation of the results:

* The User-User model provided personalised and diverse recommendations. For user 12882, the predicted scores were consistent with his preferences, with values between 3.09 and 4.09.
* The Item-Item model (ratings) showed effective groupings of popular films, but less thematic diversity.
* The content-based (genre) model ensured thematic coherence, but tended to recommend sequels or very similar films, reducing novelty.

Overall, the best performing model was the User-User Collaborative Filtering model, due to its customisability and adaptability to the user's actual tastes.

## Could Content-Based Filtering Improve Results?

Including or combining content-based filtering elements would be beneficial, especially in situations where collaborative filtering fails, such as:

* Cold start: new users or movies without sufficient data.
* Data scarcity: when there is little information shared between users.

Incorporating information such as genres, directors or actors can improve the quality of recommendations. For example, if a user has seen few films but they are all animated films, a hybrid system could suggest thematically similar films, even if there is no collaborative data available.

Therefore, a hybrid approach, combining collaborative and content-based filtering, would provide better results: more diverse, accurate and explainable.

## Interpretation of Results

Each model provided different benefits:

* The User-User model was the most effective in personalisation and showed predictions well aligned with actual user tastes.
* The Item-Item model grouped popular movies well, but without regard to subject matter.
* The content model ensured thematic consistency, but sometimes with little variety.

The overall conclusion is that while user-based collaborative filtering was the strongest model individually, the integration of content-based components can further enhance its effectiveness, especially in the face of sparse data or new users.

# Data Visualisation

## Decision of Charts

### Genre Popularity vs. Average Rating Chart

This graph shows the relationship between the popularity of movie genres (based on the total number of votes) and their average rating. A scatter plot was chosen because it allows three dimensions to be represented at once: the X-axis indicates the total number of votes (popularity), the Y-axis shows the genres, and the size and colour of the bubbles reflect the average rating. This visual approach helps to interpret how a genre can be highly voted without necessarily having good ratings.

Key decisions:

* Blue colour scale: Intuitive and neutral, where darker shades indicate better ratings.
* Maximum bubble size (40): To visually differentiate high and low rated films without cluttering the graph.
* Height of 800px: To ensure that all genres are visible without overlapping.
* Ascending order by votes: Facilitates the identification of less popular genres, preventing them from visually disappearing.

### Top 5 Movies by Genre Chart

This chart highlights the five highest rated movies within each genre, ensuring that they have at least 100 votes to avoid bias due to low ratings. A horizontal bar chart was used because titles tend to be long, and this format allows them to be readable without the need for tilting or cropping.

Key decisions:

* Vibrant blue colour (rgb(21,151,221)): Draws attention without being distracting.
* Text inside the bars: Directly displays the average rating without the need to use the cursor.
* X-axis scale (0-5): Manually set to facilitate comparisons between genres.
* Descending order by rating and votes: Ensures that the best movies appear at the top.

### Genre Trends Over Time Chart

This animated graph illustrates how the popularity of each genre has evolved over the years, highlighting trends over time. An animated bar chart per year was used, allowing the growth or decline of certain genres to be dynamically visualised.

Key decisions:

* Horizontal orientation: Favours clear reading of genre names.
* Custom colours: Each genre has a fixed colour, making it easier to recognise over time.
* 800px height: Avoids visual overlaps when there are many genres.
* Fluid animation: Allows you to intuitively explore the historical evolution.

### Top 10 Similar Movies by Genre Chart

This chart presents the ten most similar movies to a user-selected movie, based on their genre. A horizontal bar chart was used to ensure legibility of the titles and to visually represent the percentage of similarity on a well-defined axis (0-100%).

Key decisions:

* Blue colour consistent with TAB 2: Maintains a uniform aesthetic.
* Text with percentage within bars: Allows to see exact similarity without additional interaction.
* X-axis scale (0-100%): Facilitates comparison between films.
* Title cleaning: The year is removed to avoid visual confusion.
* Exclusion of selected film: Avoids artificially inflated results.

### User Behaviour Analysis Radar Chart

This chart visualises the behavioural profile of a user in five key dimensions:

1. Average rating
2. Standard deviation (variability in ratings)
3. Total number of films rated
4. Percentage of high ratings (≥4)
5. Genre diversity (measured with entropy)

A radar chart was used because it allows multiple dimensions to be represented simultaneously, providing a clear view of a user's rating ‘profile’, (ChartExpo, 2025).

Key decisions:

* Normalisation with MinMaxScaler: Ensures that all metrics are on the same scale (0 to 1), avoiding visual distortions.
* Hiding the radial axis ticks: Avoids confusion with the normalised scale.
* Hover with real values: Shows the original data for better interpretation.
* Single colour with fill: visually reinforces that it is the user's profile.
* 700px height: Prevents compression and improves legibility.

## Dashboard Design

### Interactivity and user control

The 18-35 year old audience is looking for dynamic and personalised digital experiences, (Raval, 2024). Therefore, each graph of the interface allows interaction, allowing users to choose genres, films or years. Instead of being limited to static data, tools that encourage exploration were implemented: animations in graph 3, dynamic movie selection in graph 4 and personalised analysis in graph 5. This is essential to maintain interest and provide an immersive experience, responding to the expectations of this group, who want more than just ‘looking at a chart’.

### Attractive but not overloaded visuals

Visual design is key to capturing and maintaining the attention of young users. Vibrant but professional colours were used, with a palette based on blue tones and personalisation by genre, (tenscope, 2024). Typography follows a clear hierarchy, differentiating titles, labels and content for ease of reading. In addition, animated graphics, such as the genre evolution, add dynamism and make the data ‘come alive’. This visual approach responds to the consumption habits of this generation, which is used to attractive graphics on social networks such as TikTok and Instagram, (McNamee, 2025).

### Clean and modern design

An overloaded design can demotivate users, (Akmal, 2024). Therefore, the use of flat tables and overloaded dashboards was avoided in favour of a clean, visual interface. Modern technologies such as Streamlit and Plotly were used to create a sophisticated look and feel with rounded edges, soft colours and a clear separation between sections. The integration of user-friendly interactive components helps to provide intuitive navigation, ensuring a smooth and pleasant experience.

### Fast loading times

Immediacy is paramount for this audience, who have a low tolerance for unnecessary waiting, (uptrends, 2018). With that in mind, data pre-processing was optimised so that graphics respond almost instantaneously. This optimisation ensures that users are not frustrated with long load times, improving overall usability and retention of interest in the platform.

### Neutral language and inclusive visualisation

Access to information should be intuitive, even for users with no prior knowledge of data analysis, (tenscope, 2024). Therefore, the use of unnecessary technical jargon was avoided, opting for clear and accessible language. In addition, the design was developed to be inclusive, ensuring that any user can understand the dashboard without prior experience with analytical tools.

## Justification of the data preparation process for visualisations

To analyse the data, the various datasets containing information on movies, ratings and user details were first loaded and merged. This integration was necessary because the data was spread across multiple tables, which prevented calculations such as average rating per genre without first merging the movie information with its category and rating.

Next, data cleaning was performed to improve the consistency and usefulness of the data. Rows with null or duplicate values were removed, and irrelevant columns, such as timestamp, were discarded. In addition, movie titles were cleaned by removing details such as year of release, which allowed for better grouping and avoided problems in the analysis.

Next, movie with a significant number of votes were filtered out. To ensure the validity of the metrics in charts such as ‘Top 5 by genre’, only films with at least 100 votes were considered. This was done in order to reduce bias, preventing movies with low ratings from disproportionately influencing the results.

The processing of genres was done by one-hot encoding, converting each category into a binary column. This method was essential because many movies belong to multiple genres, and a simple format such as ‘Drama|Romance’ would have made correct grouping difficult. With the proper encoding, it was possible to calculate metrics for each category.

A number of key statistical metrics were then calculated, such as the total number of votes per genres, the average rating, the user's standard deviation and the entropy of genres they have rated. The percentage of high ratings (≥4) per user was also determined. These calculations were fundamental to the visualisations, allowing us to establish patterns such as how diverse a user's taste is or how demanding they are when rating movies.

Finally, the results were sorted according to votes and ratings, depending on the type of graph. Functions such as sort\_values, nlargest and groupby were applied to structure the data logically. This allowed the best movie to appear in prominent positions and the genres with the highest number of votes to be at the end, optimising the interpretation of the results. A final formatting was also performed to improve the presentation in the interface, rounding values and generating auxiliary columns with more accessible texts, ensuring that the information was clear and attractive to users.

# Conclusion

This project demonstrates the value of data analytics and machine learning to improve movie classification and recommendation, key tools for any company managing large volumes of audiovisual content. The use of clustering allowed the segmentation of films according to their genres, where the K-Means algorithm with genre variables showed the best performance, although with room for improvement due to the multi-label nature of the data.

In recommender systems, the User-User collaborative filtering model highlighted for its customisability, although it is suggested to complement it with hybrid models that integrate content to address problems such as cold start or data sparsity.

Finally, an interactive dashboard focused on young users (18-35 years old) was developed, combining modern visuals, interactivity and clarity. These approaches can be applied in real-world contexts to provide relevant recommendations and improve the end-user experience in an effective and sustainable way.

# References

Akmal, A.M. (2024). *Less is More in User Experience Design: The Paradox of Choice*. [online] Medium. Available at: <https://medium.com/@abiyouakmal/less-is-more-in-user-experience-design-the-paradox-of-choice-82adff28e940>.

ChartExpo (2025). Radar Chart: Makes Your Complex Comparisons Simple. [online] Chartexpo.com. Available at: https://chartexpo.com/blog/radar-chart#.

McNamee, S. (2025). Social media statistics in the UK: What businesses need to know in 2025. [online] Sprout Social. Available at: <https://sproutsocial.com/insights/social-media-statistics-uk/>.

Raval, T. (2024). Council Post: Digital Transformation In The Age Of Millennials And Gen Z. *Forbes*. [online] 12 Aug. Available at: https://www.forbes.com/councils/forbestechcouncil/2019/08/20/digital-transformation-in-the-age-of-millennials-and-gen-z/.

tenscope (2024). *10 Essential Data Visualization Design Principles for Clearer Insights - Tenscope Blog*. [online] Tenscope.com. Available at: <https://www.tenscope.com/post/data-visualization-design-principles-for-clarity>.

uptrends (2018). *The psychology of web performance | The Uptrends Blog*. [online] blog.uptrends.com. Available at: https://blog.uptrends.com/web-performance/the-psychology-of-web-performance/.

# Streamlit Link

https://appca2py-bwz9dgjguscjbipn8ijpvt.streamlit.app/

# Github Link

https://github.com/AntonioGiambra/Movies\_Analysis