An Exploration of Rating Patterns in Attraction Reviews

Alyssa Garcia

Abstract— **Relying on reviews and ratings to find travel destinations that align with a frequent traveler's specific interests is often unhelpful due to the overwhelming number of options available. To simply the process of selecting travel destinations, this project aims to analyzes individual rating patterns for various attractions by identifying clusters based on rating patterns, in order to enhance travel recommendation systems. The models used are K-means clustering and K-nearest neighbors (KNN). In this project, six distinct clusters were identified and analyzed for similarities and differences. This research will allow for discovering lesser-known attractions, which will improve travel experiences and generate more business for smaller companies.**

Keywords—K-means clustering, K-nearest neighbor, travel recommendation systems, rating patterns, attraction reviews

I. Introduction

As a frequent traveler, the challenge of sorting through travel destination is overwhelming. In today's digital age, relying on reviews and ratings is a common approach in decision-making for travel. To simplify the challenge, this study examines individual rating patterns for various attractions. By identifying clusters based on these patterns, the goal is to enhance travel recommendation systems and offer more tailored travel recommendations. The main research objectives in this study revolve around exploring the distinctions in rating patterns among different clusters, analyzing the reasons behind higher rankings within specific clusters, and categorizing travelers into groups based on their rating patterns. The use of machine learning algorithms such as K-means clustering and K-nearest neighbor (KNN) are essential in exploring these topics. By using the dataset found from Kaggle [1], this research will contribute to the development of a more personalized and efficient travel recommendation system.

II. Previous Work

A. Evaluation of Partioning Clustering Algorithms for Processing Social Media Data in Tourism Domain

[2] The study conducted aimed to compare and identify suitable clustering algorithms for processing social media data in the tourism industry. Their primary objective was to explore various machine learning algorithms, including K-means, K-medoids, Clustering for Large Applications, and others, to determine which one had the better performance. Ultimately, their analysis revealed that the K-means algorithm outperformed the others.

B. Traveler's Use of Social Media: A Clustering Approach

[3] The study implemented a clustering approach to explore how travelers utilize social media platforms. The main objective of the study was to investigate the segmentation of traveler's social media content and identify distinct groups based on specific characteristics. In order to do this, the researchers used K-means algorithm and three hierarchical algorithms. The researchers identified five distinct groups of social media user, each showing different patterns in their travel-related content. The study emphasized the significance of two groups that displayed high social media creation levels, as they held the most potential in influencing others' travel decisions.

C. Understanding Tourist Behavior Towards Destination Selectioin Based on Social Media Information: An Evaluation using Unsupervised Clustering Algorithms

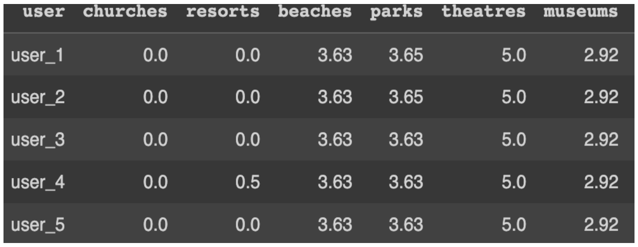
[4] The study aimed to gain a deeper understanding of tourist behavior regarding destination selection through the analysis of social media. In order to do this, the researchers applied clustering algorithms, including hierarchical clustering, Expectation-Maximization, and model-based clustering. The researchers identified eight distinct groups of tourists across East Asia. These clusters displayed high levels of attraction towards parks and picnic spots, indicating popular destinations among the tourists. The study revealed that some clusters shared strong similarities, indicating common preferences and interest for certain groups of tourists, while others showed significant differences in their destination choices.

III. Methodology

A. Dataset

The dataset used in this project was sourced from Kaggle [1] and consists of 5457 rows of user ratings. Each rating is measures on a scale from zero to five, reflecting an individual user. The dataset contains 25 columns, each representing a different attraction including churches, resorts, beaches, parks, theatres, museums, malls, zoos, restaurants, pubs and bars, local services, burger and pizza shops, hotels and other lodgings, juice bars, art galleries, dance clubs, swimming pools, gyms, bakeries, beauty and spas, cafes, viewpoints, monuments, and gardens. Some of these attractions can be seen in Table I. These columns contain the users' preferences towards various attractions, providing valuable insights to their interests.

TABLE 1. DISPLAYS SOME ATTRACTION COLUMNS



user	churches	resorts	beaches	parks	theatres	museums
user_1	0.0	0.0	3.63	3.65	5.0	2.92
user_2	0.0	0.0	3.63	3.65	5.0	2.92
user_3	0.0	0.0	3.63	3.63	5.0	2.92
user_4	0.0	0.5	3.63	3.63	5.0	2.92
user_5	0.0	0.0	3.63	3.63	5.0	2.92

B. *Data Cleaning*

For data processing and cleaning, Python along with the pandas library was used. To ensure readability, all columns were renamed to their respective attraction. One challenge encountered during this process was the column “local”, was classified as an object type. To fix this, the column was converted to the appropriate data type, float64. To eliminate any extra users, the dataset was reviewed for duplicate users. The missing values were addressed by removing the miscellaneous column, which was missing 5454 values. To address the missing values in the gardens and burger_pizza columns, the median values were used. In addition, extreme outliers were identified and removed from the dataset. This process built a foundation for a reliable analysis.

C. *Models*

- K-Mean Clustering – An unsupervised machine learning algorithm that sorts data points into K clusters based on their similarities. The “K” in K-means represents the number of clusters inputted into algorithm.

- KNN – A supervised machine learning algorithm that is used for both classification and regression. KNN predicts or classifies data points by considering the closets neighbors to that point. It assumes the data points with similar characteristics belong to the same group or have similar values.

D. *Performance Metrics*

Performance metrics are used to access the accuracy of each machine learning model. In the testing and training phases, the following performance metrics were applied:

- Silhouette Score – Calculates how well-separated the clusters are from each other.

$$Silhouette\ Score = \frac{(b(i)-a(i)}{\max(a(i),b(i))} \qquad (1)$$

a = average distance to the points in the nearest cluster
b = average intra-cluster distance to all the points

- Accuracy – Measures proportion of correct predictions out of total predictions.

$$Accuracy = \frac{Correct\ Predictions}{Total\ Number\ of\ Predictions} \qquad (2)$$

- K-Fold Cross Validation – Evaluates prediction ability of the model.

$$Average = \frac{1}{k} E_{i=1}^{k} x_i \qquad (3)$$

x_i = value obtained evaluating the model on test data from fold i.

IV. Results

A. *K-Means Clustering*

The K-means algorithm was used to analyze the rating patterns of individual travelers and categorize them into groups based on their similarities. Each cluster represents a group of travelers with similar preferences for the 24 attraction categories. The optimal number of clusters identified in this project was six clusters, which was determined using the Silhouette analysis with a score of 0.15, which can be seen in Figure 1. This score indicates that the clusters are separated but may overlap, indicating some similarities between the groups. Each cluster represents distinct preferences of the users, which can also be seen in Figure 2:

Cluster 0 represents individuals who consistently gave high ratings to malls, restaurants, and art galleries. These individuals are more likely to enjoy shopping, dining, and exploring art exhibits during their trip.

Cluster 1 represents individuals who consistently gave high ratings to parks, theaters, and viewpoints. These individuals are likely to enjoy outdoor activities and cultural experiences.

Cluster 2 represents individuals who consistently gave high ratings to gardens, monuments, viewpoints, and bakeries. These individuals are likely to enjoy historical landmarks, scenic views, and baked goods.

Cluster 3 represents individuals who consistently gave high ratings to restaurants, pubs, and local services. These individuals are likely to enjoy local dining and services during their trips.

Cluster 4 represents individuals who consistently gave high ratings to burger and pizza shops, hotels, and juice bars. These individuals are likely to enjoy fast-food options and hotels for their trips.

Cluster 5 represents individuals who consistently gave high ratings to theaters, museums, and malls. These individuals are likely to enjoy cultural experiences and shopping.

In terms of individual travelers, Table 2 illustrates the difference in rating patterns for two specific travelers. User_1 was assigned to Cluster 5, which indicates their preference for theatres, museums, and malls, while User_79 was grouped into Cluster 0, reflecting their higher ratings for malls, restaurants, and art galleries.

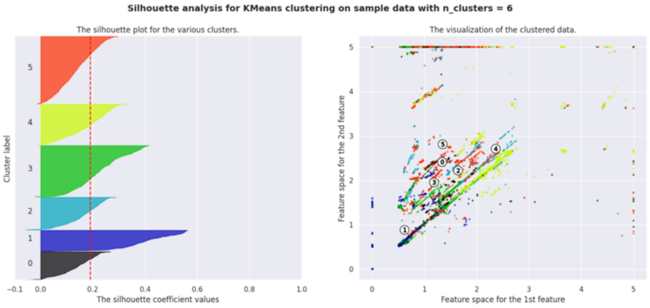


FIGURE 1. SILHOUETTE ANALYSIS FOR 6 CLUSTERS

Review distribution of Cluster#0
Review distribution of Cluster#1
Review distribution of Cluster#2
Review distribution of Cluster#3
Review distribution of Cluster#4
Review distribution of Cluster#5

FIGURE 2. DISPLAYS THE CLUSTER DISTRIBUTIONS

TABLE I. ILLUSTRATES THE DIFFERENCE BETWEEN USERS

User	Cluster	Theatre Rating	Museum Rating	Mall Rating	Art Galleries Rating	Restaurant Ratings
1	5	5.0	2.92	5.0	1.74	2.33
79	0	3.92	5.0	2.88	1.17	2.71

B. *K-nearest neighbor*

The results of the KNN algorithm displayed a proficiency in accurately predicting cluster assignments for new datapoints, achieving an accuracy performance of 95%. This level of accuracy indicates the overall effectiveness of KNN in classifying individuals into their respective cluster.

In addition to using accuracy to evaluate the KNN model, K-Fold Cross Validation was implemented for further analysis. The K-Fold Cross Validation produced a mean accuracy of 0.94, which ensured the accuracy of the KNN algorithm. The confusion matrix, which can be seen in Figure 3, displays the results of the K-Fold Cross Validation by providing a prediction summary for each fold.

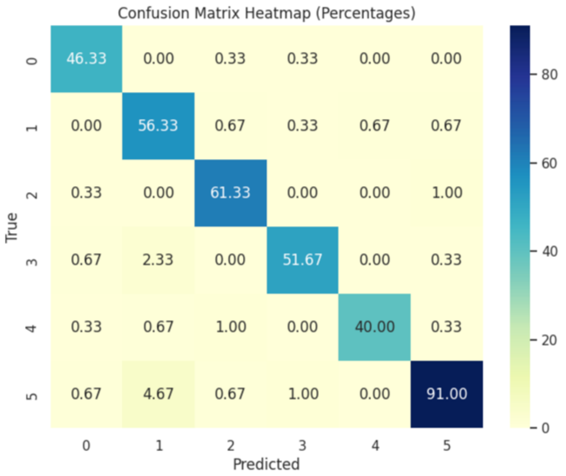


FIGURE 3. CONFUSION MATRIX HEATMAP

V. CONCLUSION

This project utilized K-means clustering and K-nearest neighbor (KNN) algorithms to analyze individual rating patterns and categorize travelers into distinct clusters based on their preferences for various attractions. The K-means algorithm successfully identified six clusters; each represent a group of travelers with similar interests. Additionally, the KNN algorithm demonstrated proficiency in accurately predicting cluster assignments for new data points, achieving an accuracy of 95%. This project has potential to enhance travel recommendation systems, generate more business for smaller companies, and provide traveler's with tailored suggestions that align with their preferences.

REFERENCES

[1] Kaggle.(n.d.) Travel Review Rating Dataset. Retrieved from https://www.kaggle.com/datasets/wirachleelakiatiwong/travel-review-rating-dataset

[2] Renjith, D., Sreekumar, A., & Jathavedan, M. (2018). Evaluation of Partitioning Clustering Algorithms for Processing Social Media Data in Tourism Domain. *2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, 127-131

[3] Amaro, Duarte, P., & Henriques, C. (2016). Travelers' use of social media: A clustering approach. Annals of Tourism Research, 59, 1–15. https://doi.org/10.1016/j.annals.2016.03.007

[4] Ghosh, & Mukherjee, S. (2023). Understanding tourist behaviour towards destination selection based on social media information: an evaluation using unsupervised clustering algorithms. JOURNAL OF HOSPITALITY AND TOURISM INSIGHTS, 6(2), 754–778. https://doi.org/10.1108/JHTI-11-2021-0317