Title: Classification Holiday Sports Destinations in South India

Student ID: s4075995

Student Name and email (contact info): Ramneek Kaur Riar s4075995@student.rmit.edu.au

Affiliations: RMIT University.

Date of Report: 29th May 2024

I certify that this is all my own original work. If I took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in my submission. I will show I agree to this honor code by typing "Yes": *Yes*.

1 Abstract 3
2 Introduction 3
3 Methodology 3
4 Results 4
5 Discussion 11
6 Conclusion 12
References 12

1 Abstract

This report aims to classify a user's interest in sports-based attractions in South India based on their preferences in other recreational activities that they have reviewed. The dataset comprises of 249 users, with attributes (Sports, Religious, Nature, Theatre, Shopping, Picnic) containing the number of reviews they have given. The sports attribute was chosen to be the target class. Users were categorized as either "High" or "Low" based on the number of reviews they gave to the sports category. The data set was pre-processed and explored to understand the attributes at a better level and derive relationships amongst individual attribute pairs. Decision Tree and K-Nearest Neighbour classifiers were then trained with this data with various training and testing split levels. This helped in determining which model with which training testing split provided results with the highest accuracy in predicting interest in sport-based attractions in South India per user. Findings determine that interest in certain attractions than others can determine a user's interest in sports-based attractions. This can help the tourism industry in South India to target sports-based attractions to certain individuals and make personalized recommendations.

2 Introduction

Understanding interest levels in particular tourist activities can foster growth in the tourism industry within that country. This report explores the relationship between users' interests in recreational activities in South India and their interest in sport-related activities. The dataset used for this research consists of 249 individual users with their recorded number of reviews per category which are: Religious, Nature, Theatre, Shopping, Picnic and Sports. With this report, the individual destination categories are explored on their own and with one another to discern descriptive insights using graphs. Then, data modelling is performed using two models (Decision Trees and K-Nearest Neighbours) which help determine either high or low interest levels for sports-based activities based on data already acquired. This can help make recommendations to tourists in South India on whether sports-based activities will be suitable for them to explore.

3 Methodology

3.1 Data Set

The dataset used in this report was obtained from UC Irvine Machine Learning Repository at https://archive.ics.uci.edu/dataset/476/buddymove+data+set (Renjith, 2018). The data set contains 249 records with 7 attributes, including the target attribute and they are:

- User Id: Unique user Id of the user (ignored for the sake of exploration and modelling)
- Sports: Number of reviews by user on sports-based tourist attractions **(target class)**
- Religious: Number of reviews by user on religious-based tourist attractions
- Shopping: Number of reviews by user on shopping-based tourist attractions
- Theatre: Number of reviews by user on theatre-based tourist attractions
- Picnic: Number of reviews by user on picnic-based tourist attractions
- Nature: Number of reviews by user on nature-based tourist attractions

3.2 Classification Techniques

The sports category was made into the target class for the sake of this classification problem, determining whether the user would have low or high interest levels in exploring sports-based attractions. To do this, binarization was performed on the sports category data by finding out the median value within the column. The median was 12, so the number of reviews less than the median were given a value of 0 (0 = low interest levels) and the number of reviews higher than the median were given a value of 1 (1 = high interest levels).

The following classification techniques were used to classify the target class: Decision Trees and K-Neatest Neighbours. The dataset was split into training and test ratios of different sizes for these classification methods (60/40, 50/50, and 80/20) to help determine which depth value or k-value provided the optimal accuracy for the model, before the model starts to overfit or underfit the data. The classification error rate was also calculated, as well as precision, recall and f1-score obtained from the model's classification report to further help measure the model's accuracy.

4 Results

4.1 Retrieving and Preparing the Data

The goal of the project was to successfully classify whether a user would have low or high interest in visiting sports-based attractions in South India, based on the number of reviews they have given to other attraction categories. To begin with data preparation, verification was performed to confirm whether the loaded data was equivalent to the data in the source CSV file. This was done by loading the data twice into separate data frames and then compared with one another to ensure the data was consistent. After data types were examined to ensure that all numerical attributes contained numerical data types which was successful as all columns contained the type of int64. Then, the data was examined to confirm that no NaN (Not a Number) values existed in the data. The sum of NaN values was 0. Lastly, the data was checked to drop any duplicate data that may exist within the file. This means that any rows which contain identical information, such as the same user Id with the same number of reviews per category. After the data was cleaned, the User Id column was dropped as it had no significance in retrieving results for data exploration or classification. The clean data set was saved into the file "buddymove_holidayiq_clean.csv" and was ready to be used for exploration and analysis.

4.2 Data Exploration Per Attribute

The clean data was retrieved and the total number of reviews per category was calculated and presented in a pie chart. As shown in Figure 4.1, the sports category had the least number of reviews at 2%, while the rest of the categories had consistent percentages throughout between a range of 18% to 21%. The nature category had the highest number of reviews (31005) while the sports category had the least number of reviews (2985). There is also a visible pattern in the categories with travellers preferring outdoor activities the most (Picnic and Nature), then entertainment/recreational activities (Theatre, Shopping and Religious) and lastly physical activity (Sports).

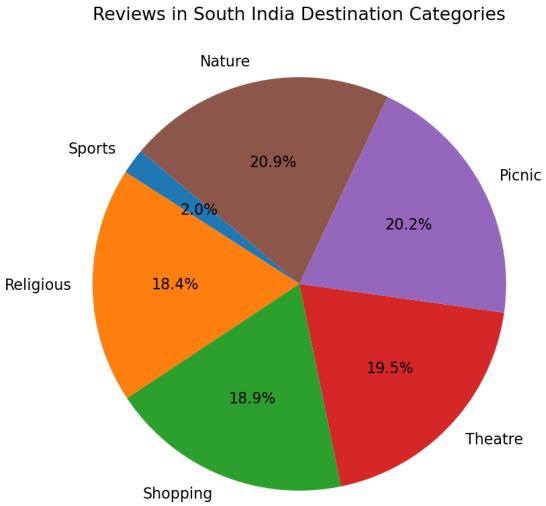


Figure 4.1: Total Sum of Reviews in South India Destination Categories

4.2.1 Sports

The density plot of the Sports attribute is shown in Figure 4.2 below, which shows that most users visiting South India went to around 5-8 sports-based attractions as it as the highest density. The next peak in density happens around 20 reviews, which means these users display a high interest in visiting sports-based attractions.

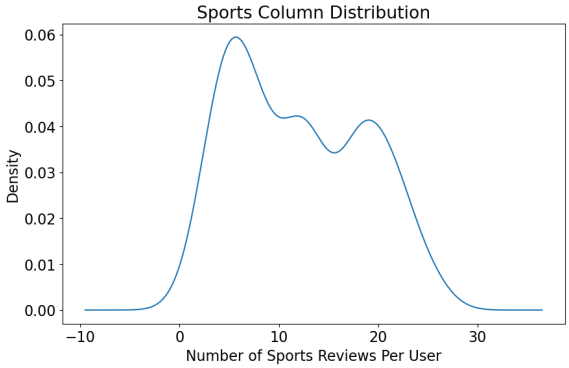


Figure 4.2: Sports Column Distribution

4.2.2 Religious

The histogram in Figure 4.3 shows the frequency of user reviews in the religious category. Bin 3 (with range 80 to 100) shows the highest number of reviews at 58, whereafter the graph peaks and continues in a downward trend. The bulk of the revies land between the range of 80 to 140 showing that religious sites are quite popular amongst tourists in South India.

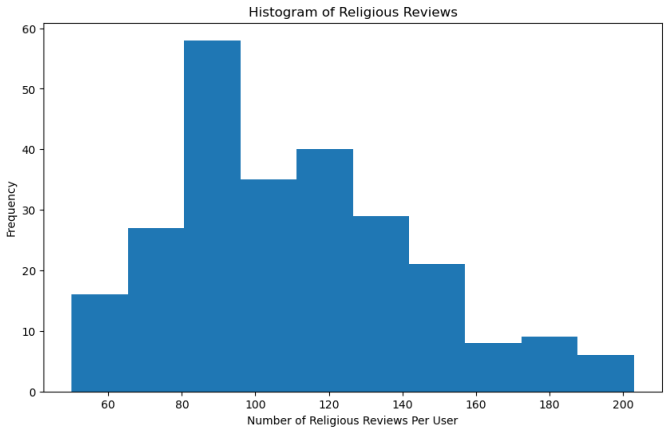


Figure 4.3: Religious Column Distribution

4.2.3 Shopping

The histogram in Figure 4.4 below shows the frequency of user reviews in the shopping category. Bin 3 (with range 85 to 105) shows the highest number of reviews at 49, whereas the bin right after shows competitive levels alongside the rest of the graph. Overall shopping is a popular tourist attraction as frequency is relatively high even as the number of reviews increase.

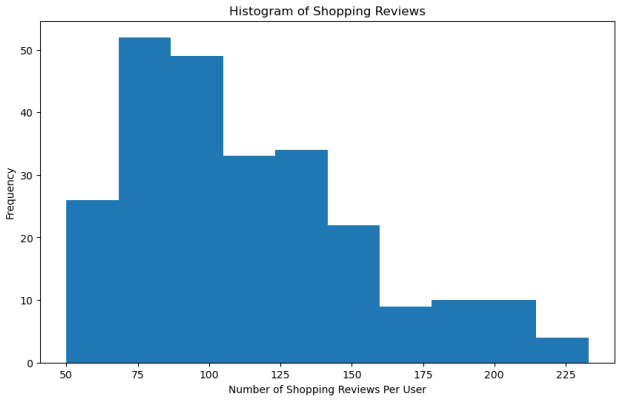


Figure 4.4: Shopping Column Distribution

4.2.4 Theatre

The histogram in Figure 4.5 below shows the frequency of user reviews in the theatre category. Bin 4 (with range 105 to 121) shows the highest number of reviews at 49, and the graph shows a steady decline in frequency as the number of reviews increase from a specific user. This means that bulk of the users prefer to visit around 60-121 theatre-based attractions, with more reviews after showing a niche-based interest.

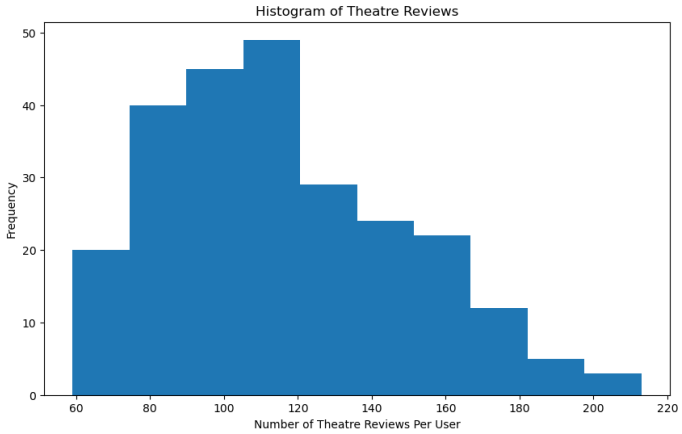


Figure 4.5: Theatre Column Distribution

4.2.5 Picnic

The histogram in Figure 4.6 shows the frequency of user reviews in the picnic category. Bin 2 (with range 76 to 93) shows the highest number of reviews at 48, where the graphs then dips and hits another peak in bin 5 (with range 123 to 140) with 42 reviews. This shows that picnic-based attractions are relatively popular with the greatest number of reviews by a user sitting in the 76 to 140 range.

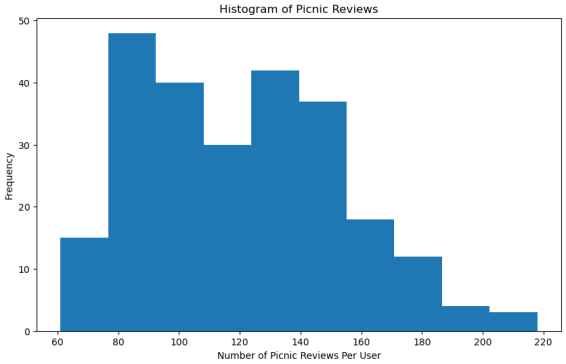


Figure 4.6: Picnic Column Distribution

4.2.6 Nature

The histogram in Figure 4.7 shows the frequency of user reviews in the nature category. Bin 3 (with range 105 to 132) shows the highest number of reviews at 50, and the graph shows a steady decline in frequency as the number of reviews increase from a specific user. This means that most if not all users tend to visit many nature-based attractions in South India in this range, whereas the steady downwards trend shows a more niche-based interest.

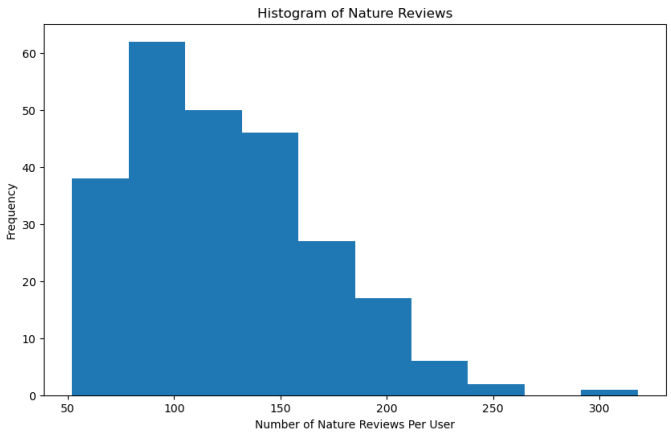


Figure 4.7: Nature Column Distribution

4.3 Data Exploration Per Attribute Relationship

In total there are 15 pair relationships determined from the individual columns, however the 10 relationships focused on in this report are the following explained below.

Sports – Nature

Figure 4.8 displays a scatterplot of the number of sports reviews per user and the number of nature reviews per user. As the number of sports reviews increase, so do the number of nature reviews for that user. This shows a strong correlation between nature and sports-based attractions, often because these can be based in an outdoor setting.

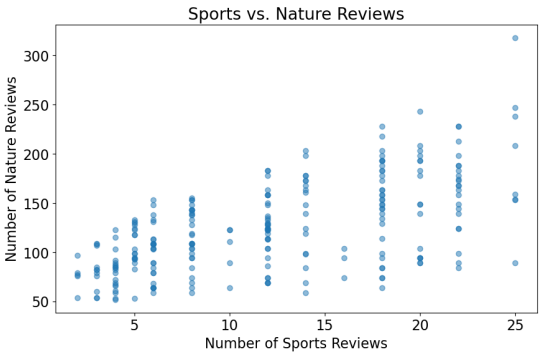


Figure 4.8: Scatterplot of Sports vs. Nature Reviews Per User

Sports – Theatre

The relationship between the number of reviews in the sports and theatre category per user was graphed in a scatterplot. Overall, the graph shows an upward trend, displaying that as the number of sports reviews increase so do the theatre reviews.

Sports – Shopping

The relationship between the number of reviews in sports and shopping category per user was graphed in a

scatterplot. The graph shows a steeper upward trend compared to the previous graphs, showing an increased likelihood of shopping reviews as the number of sports reviews grow. This can be due to many shopping centres having proximity to other tourist attractions which would explain the steep rise.

Sports – Picnic

Figure 4.9 shows the relationship between the number of reviews in sports and picnic category per user in a scatterplot. The graph shows a clear upward trend, showing an increased likelihood of picnic reviews as the number of sports reviews grow. This can be due to many picnic areas having proximity to other areas where sports activities are highly prevalent, such as parks and other outdoor regions.

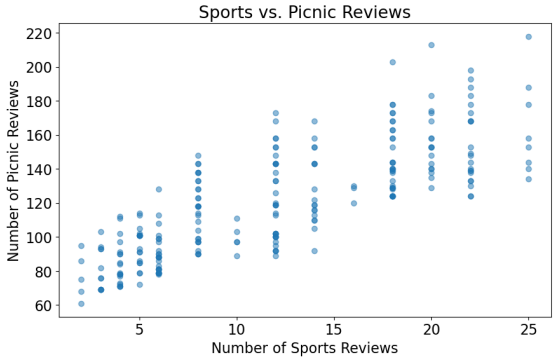


Figure 4.9: Scatterplot of Sports vs. Picnic Reviews Per User

Religious – Nature

The relationship between the number of reviews in religious and nature category per user was graphed in a scatterplot. The graph shows an almost quadratic relationship between these variables, with the number of religious reviews at 120 has the highest number of nature reviews at around 230 reviews, before the number of nature reviews starts to decrease. This suggests that religious attractions are often built around nature attractions, and as users start to decrease their visits to religious sites, so do their visits to nature sites decrease.

Religious – Picnic

Figure 4.10 shows the relationship between the number of reviews in religious and the picnic category per user in a scatterplot. The graph shows a clear upward trend, showing an increased likelihood of religious reviews as the number of picnic reviews grow. This can be due to many picnic areas having proximity to other areas where religious sites are highly prevalent, such as outdoor regions.

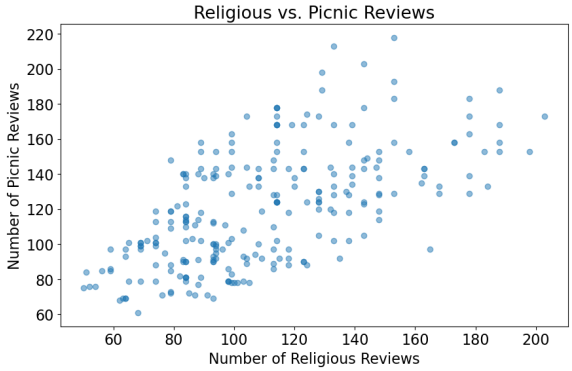


Figure 4.10: Scatterplot of Religious vs. Picnic Reviews Per User

Nature – Theatre

Figure 4.11 shows the relationship between the number of reviews in nature and the theatre category per user in a scatterplot. The graph shows a clear upward trend, showing an increased likelihood of nature reviews as the number of theatre reviews grow. However, this seems to hit a limit at around 200 nature reviews whereafter the graph starts to fall off. Outliers suggest that reviewers who visit more theatre-based attractions are less likely to visit nature-based attractions and vice-versa.

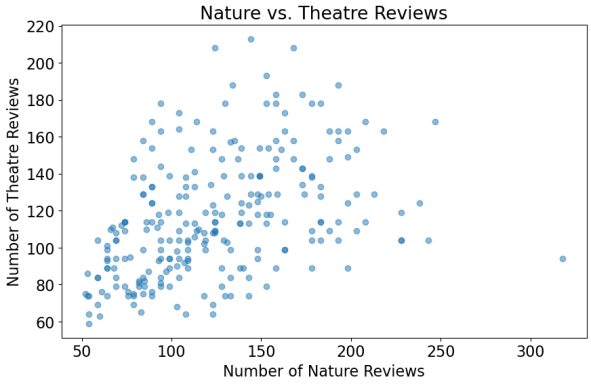


Figure 4.11: Scatterplot of Nature vs. Theatre Reviews Per User

Nature – Picnic

Figure 4.12 shows the relationship between the number of reviews in nature and the picnic category per user in a scatterplot. The graph shows a clear upward trend, showing an increased likelihood of nature reviews as the number of picnic reviews grow. However, this seems to hit a limit at around 250 nature reviews whereafter the graph starts to fall off. This is due to there not being enough data within our data set.

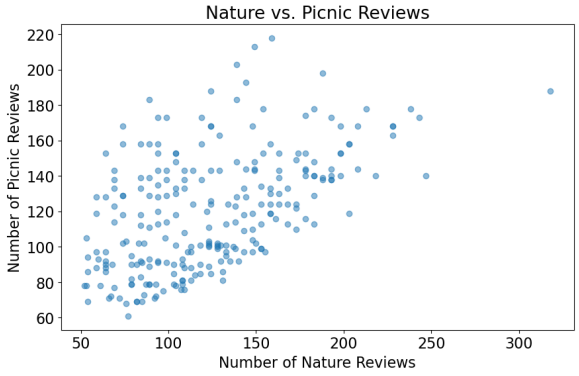


Figure 4.12: Scatterplot of Nature vs. Picnic Reviews Per User

Theatre – Shopping

The relationship between the number of reviews in theatre and shopping category per user was graphed in a scatterplot. The graph shows an upward trend when given a line of best fit, showing an increased likelihood of theatre reviews as the number of shopping reviews grow. This can be since both theatre and shopping are mostly indoor activities, so users who prefer one activity will likely have the same preference for the other.

Theatre – Picnic

The relationship between the number of reviews in theatre and picnic category per user was graphed in a scatterplot. The graph shows a steady upward trend compared to the previous graphs, showing a likelihood of theatre reviews as the number of picnic reviews grow. This can be due to many shopping centres having proximity to other tourist attractions which would explain the steep rise.

4.4 Data Modelling

4.4.1 Decision Tree Classifiers

Figure 4.13 shows the process to find the optimal tree depth for each training/testing split, whereafter in the graph the classifier accuracy decreases due to model overfitting or underfitting. For the 60/40 classifier the optimal tree depth is 5, and for the 50/50 and 80/20 split it is 4.

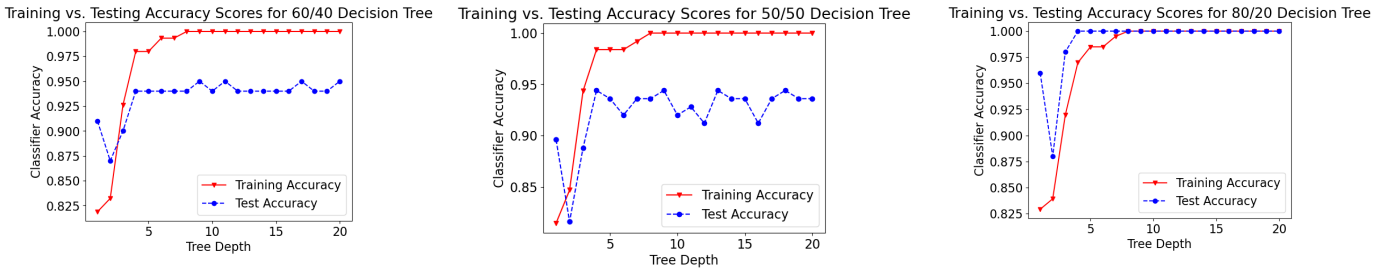


Figure 4.13: Training vs Testing Accuracy Scores for Various DT Split Levels

	Accuracy (%)	Classification Error Rate (%)	Confusion Matrix
60/40 Split	93.0	7.0	[40 5] [2 53]
50/50 Split	93.60000000000001	6.4	[52 5] [3 65]
80/20 Split	100.0	0.0	[23 0] [0 27]

Table 1: Decision Tree Accuracy, Classification Error Rate and Confusion Matrix per Split

Table 1 above shows the accuracy, classification error rate and confusion matrix for each training and testing split. The 80/20 split showed the highest accuracy at 100%, however this can be skewed as the number of records were already limited at 249. If there had been more instances, it could have reflected an even accurate representation of the model. Figure 4.14 below shows the precision, recall and f-1 scores for each split as well.

Classification Report:

	precision	recall	f1-score	support
0	0.95	0.89	0.92	45
1	0.91	0.96	0.94	55
accuracy			0.93	100
macro avg	0.93	0.93	0.93	100
weighted avg	0.93	0.93	0.93	100

Classification Report:

	precision	recall	f1-score	support
0	0.95	0.91	0.93	57
1	0.93	0.96	0.94	68
accuracy			0.94	125
macro avg	0.94	0.93	0.94	125
weighted avg	0.94	0.94	0.94	125

Classification Report:

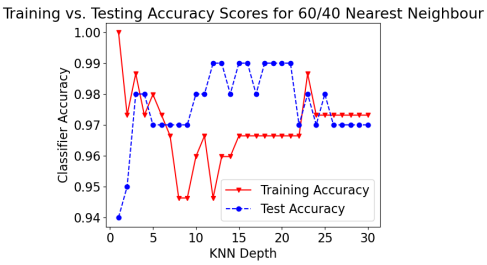
	precision	recall	f1-score	support
0	1.00	1.00	1.00	23
1	1.00	1.00	1.00	27
accuracy			1.00	50
macro avg	1.00	1.00	1.00	50
weighted avg	1.00	1.00	1.00	50

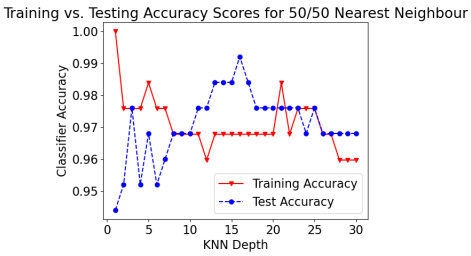
Figure 4.14:

Classification Report for each Decision Tree Split in Table Order

4.4.1 K-Nearest Neighbour

Figure 4.15 shows the process to find the optimal k value for each training/testing split, whereafter in the graph the classifier accuracy decreases due to model overfitting or underfitting. For the 60/40 and 50/50 classifiers the optimal k value is 3, while for the 80/20 classifier the k value is 7.





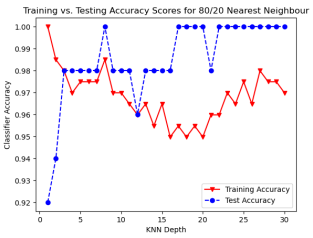


Figure 4.15: Training vs Testing Accuracy Scores for Various KNN Split Levels

	Accuracy (%)	Classification Error Rate (%)	Confusion Matrix
60/40 Split	98.0	2.0	[43 2] [0 55]
50/50 Split	97.6	2.4	[55 2] [1 67]
80/20 Split	98.0	2.0	[22 1] [0 27]

Table 2: KNN Accuracy, Classification Error Rate and Confusion Matrix per Split

Table 2 above shows the accuracy, classification error rate and confusion matrix for each training and testing split. The 60/40 and 80/20 split showed the highest accuracy at 98%, however this can be skewed as the number of records were already limited at 249. If there had been more instances, it could have reflected an even accurate representation of the model. Between these two models, it would be better to go with the 80/20 split as it allocates more training data for the model. Figure 4.16 below shows the precision, recall and f-1 scores for each split as well.

Classification Report:

	precision	recall	f1-score	support
0	1.00	0.96	0.98	45
1	0.96	1.00	0.98	55
accuracy			0.98	100
macro avg	0.98	0.98	0.98	100
weighted avg	0.98	0.98	0.98	100

Classification Report:

	precision	recall	f1-score	support
0	0.98	0.96	0.97	57
1	0.97	0.99	0.98	68
accuracy			0.98	125
macro avg	0.98	0.98	0.98	125
weighted avg	0.98	0.98	0.98	125

Classification Report:

	precision	recall	f1-score	support
0	1.00	0.96	0.98	23
1	0.96	1.00	0.98	27
accuracy			0.98	50
macro avg	0.98	0.98	0.98	50
weighted avg	0.98	0.98	0.98	50

Figure 4.16: Classification Report for each KNN Split in Table Order

5 Discussion

Figure 4.17 shows the 6 different classifiers (between Decision Trees and K-Nearest Neighbours) in a bar graph displaying their different accuracy levels. Through this, it can be observed that the Decision Tree Classifier with the 80/20 split provides the most optimal results in helping determine whether a user has high or low levels of interest in sports-based attractions in South India, dependent on their other reviews. It was seen that attractions/activities that are most likely based outdoors have a higher correlation to other attractions/activities that are outdoors since it allows tourists for easy access. The same can be said for the relationship between indoor attractions/activities. Users who have higher preference for visiting nature and religious sites have a high interest in sports-based attractions, meanwhile users who have a higher preference for theatre and shopping have a low interest in sports-based attractions.

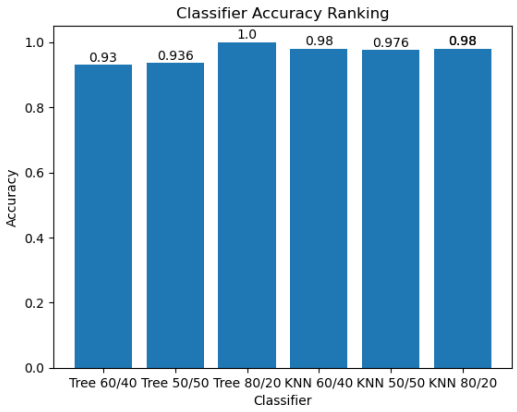


Figure 4.17: Classifier Accuracy Rankings

6 Conclusion

By leveraging the data set comprising of 249 records and 7 attributes, their interest levels in sports-based attractions were predicted. This analysis helped determine that users already showing higher levels of interest in activities that took place outdoors (Nature and Religious) were more likely to display high levels of interest in attractions falling under the sports category. On the other hand, users who showed higher interest in indoor based activities (Shopping and Theatre) were more likely to display low levels of interest in attractions falling under the sports category. These results obtained make sense as they align with the relative behaviour tourists who often plan their trips based on their interests and attractions in close proximity to one another to maximize their time spent. Classification models were created using the Decision Tree and K-Nearest Neighbour algorithm, however the Decision Tree model with an 80/20 training and test split with a tree depth of 4 obtained the highest accuracy. These results can be helpful to the tourism industry in South India as they continue to grow tourism within the region, offering personalized recommendations to tourists to increase their enjoyment whilst simultaneously fostering positive reviews for their attractions.

References

Renjith, Shini. (2018). “Buddymove Data Set” UCI Machine Learning Repository, https://archive.ics.uci.edu/dataset/476/buddymove+data+set