Title: Comparative Analysis of KNN and Decision Tree Classifiers for Predicting User Interest in Sports

Student ID: **s3960736**
Student Name): **Shaikh Mohammad Rahil**
and Email (Contact Info s3960736**@student.rmit.edu.au**
Affiliations: **RMIT University**
Date of Report: **29 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.**

Contents

Abstract

This report presents a comprehensive analysis of predicting user interest in sports using the BuddyMove Data Set. This dataset, sourced from user reviews on holidayiq.com, contains counts of reviews across six categories of destinations in South India. The study involves data preparation, exploration, and modelling to determine the most effective classification model between K-Nearest Neighbors (KNN) and Decision Tree classifiers.

The data preparation phase included handling missing values, encoding categorical variables, and standardizing the features for consistent scaling. During data exploration, descriptive statistics and visualizations were employed to uncover underlying patterns and correlations within the dataset. For instance, a strong positive correlation was observed between 'Nature' and 'Picnic' reviews, indicating that users interested in nature-related activities are also likely to engage in picnics. Conversely, a negative correlation between 'Religious' and 'Theatre' reviews suggested differing user preferences for these categories.

In the data modelling phase, both KNN and Decision Tree classifiers were trained and evaluated using several metrics such as accuracy, precision, recall, F1-score, and cross-validation accuracy. The KNN classifier was optimized with weights='distance' and p=1, resulting in an optimal number of 6 neighbors. The Decision Tree classifier, on the other hand, was fine-tuned through feature selection techniques to improve its predictive accuracy.

The results indicated that the KNN classifier outperformed the Decision Tree classifier across all evaluation metrics. Specifically, the KNN model achieved an accuracy of 0.9600, a precision of 0.9607, a recall of 0.9600, and an F1-score of 0.9602. The cross-validation accuracy for the KNN model was 0.9800, highlighting its robustness and reliability. In comparison, the Decision Tree classifier attained an accuracy of 0.9467, a precision of 0.9467, a recall of 0.9467, and an F1-score of 0.9467, with a cross-validation accuracy of 0.9398.

Based on these findings, the KNN classifier is recommended for predicting user interest in sports due to its superior performance and consistency. The report concludes with a discussion on the implications of these results and suggestions for future work, including the exploration of additional classification algorithms and feature engineering techniques to further enhance model performance.

This study demonstrates the importance of thorough data preparation and exploration, as well as the need for rigorous model evaluation to ensure accurate and reliable predictions.

Introduction

The objective of this project is to predict user interest in sports using the BuddyMove Data Set. The dataset, sourced from user reviews on holidayiq.com, includes counts of reviews in six categories of destinations across South India. This project focuses on a classification task, comparing the performance of KNN and Decision Tree classifiers to determine the most accurate model.

Methodology

Retrieving and Preparing the Data

The BuddyMove Data Set was retrieved from the UCI Machine Learning Repository and converted to CSV format for compatibility with Python data analysis tools. The dataset includes user reviews in six categories of destinations across South India. The following features were included in the dataset:

- User Id
- Religious
- Nature
- Theatre
- Shopping
- Picnic
- Sports

The goal of this project is to predict the 'Sports' category based on user reviews in other categories. During data preparation, missing values were handled, and features were standardized to ensure consistent scaling.

Data Preparation Steps:

1. **Handling Missing Values:** Ensured there were no missing values in the dataset, as missing values can lead to inaccurate models.

2. **Standardizing Features:** Applied standardization to the features to bring all values to a common scale, which is essential for distance-based algorithms like KNN.

Data Exploration

Data exploration involved descriptive statistics and visualizations to uncover patterns and correlations within the dataset.

Descriptive Statistics:

- Summary statistics for each feature, such as mean, median, standard deviation, and range, were calculated to understand the distribution of data.

Visualizations:

1. **Histograms:** Provided insights into the distribution of each feature.

2. **Box Plots:** Highlighted the presence of outliers and the spread of data.

3. **Correlation Matrix:** Identified the strength and direction of relationships between pairs of features.
4. **Scatter Plots:** Showed relationships between pairs of numerical features.
5. **Bar Plots:** Displayed the count of reviews in each category.

Example Visualizations:

Figure 1: Histogram of Religious, Sports and Nature Reviews

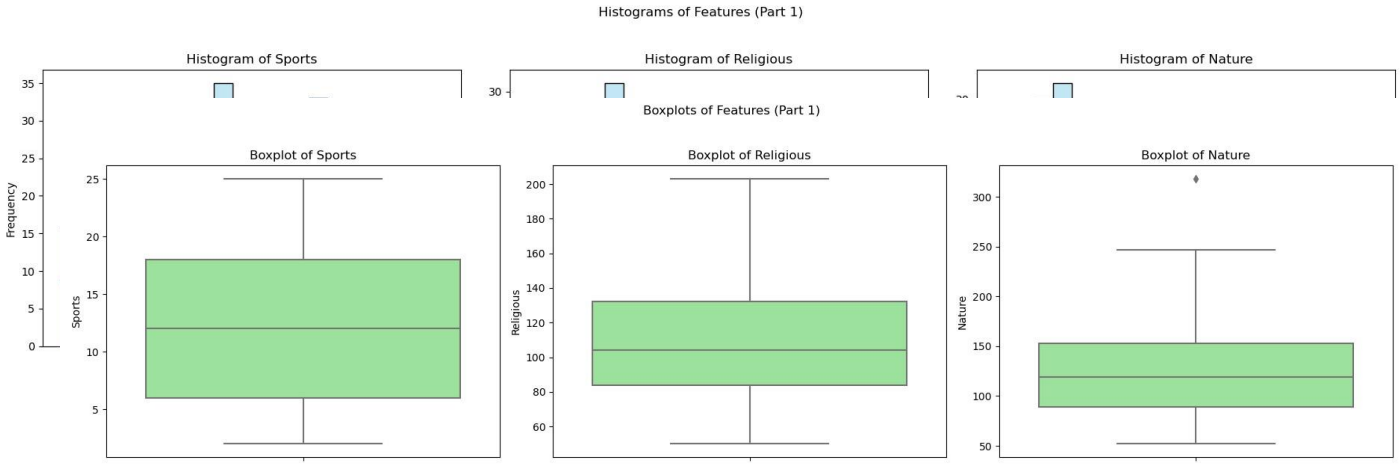


Figure 2: Box Plot of Religious, Sports and Nature Reviews

Figure 2: Scatter plot of attribute pairs with pearson coefficients.

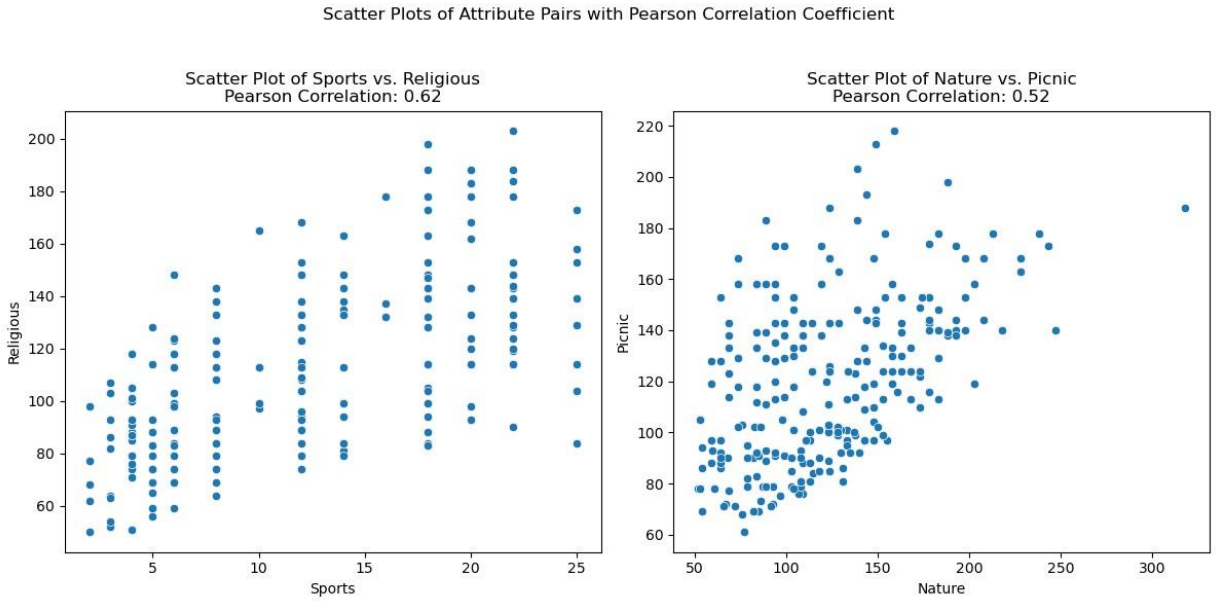
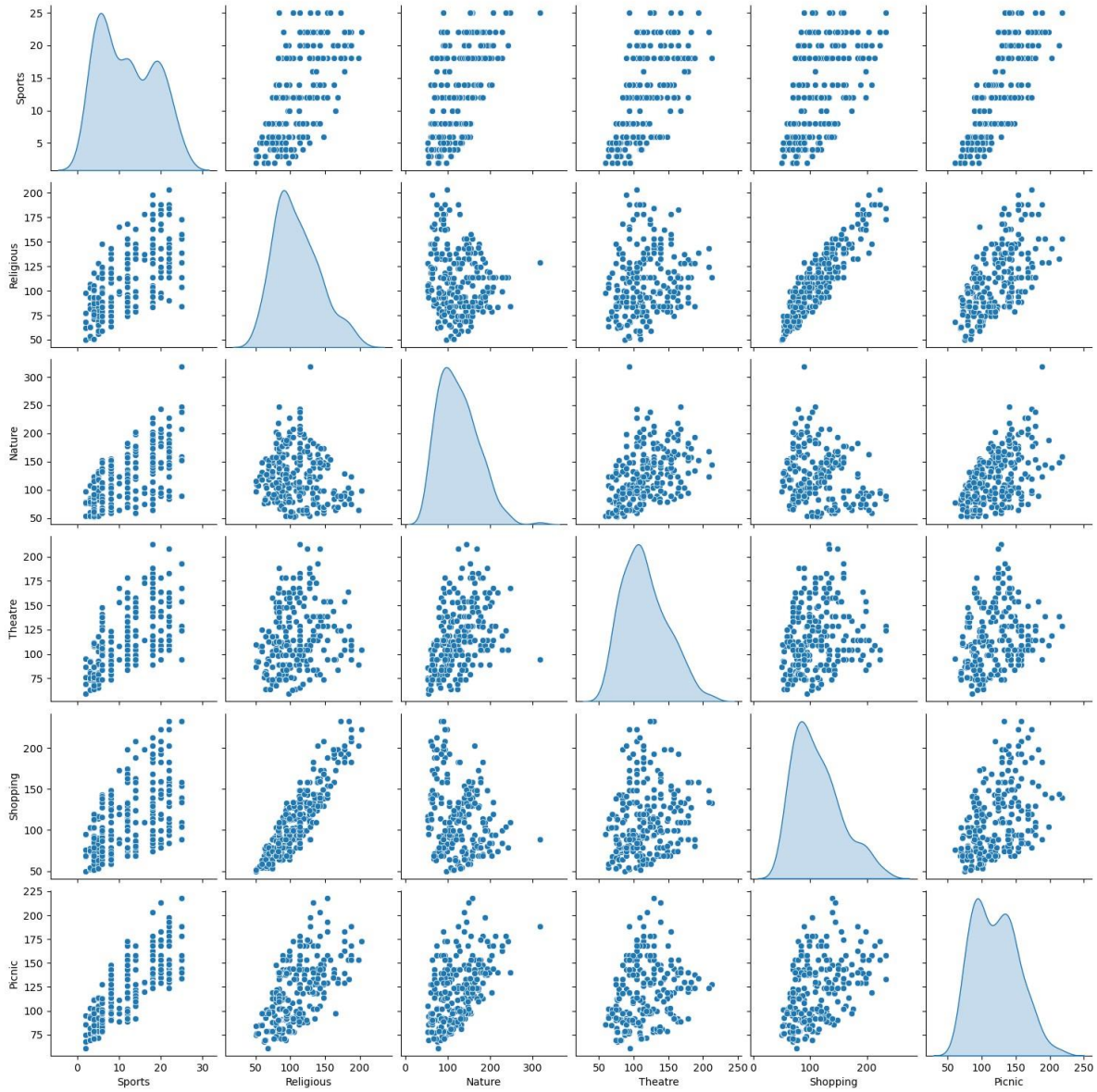


Figure 3: Scatter Plot of Nature vs. Picnic Reviews



From the data exploration, several notable patterns were identified:

- A strong positive correlation between 'Nature' and 'Picnic' reviews suggested that users interested in nature-related activities are also likely to engage in picnics.
- A negative correlation between 'Religious' and 'Theatre' reviews indicated differing user preferences for these categories.

Data Modelling

Two classification models, KNN and Decision Tree, were developed and evaluated to predict user interest in sports.

Model Building Steps:

1. **Feature Selection:** Relevant features were selected based on their correlation with the target variable and their contribution to model accuracy.

2. **Training with Default Parameters:** Both models were initially trained using default parameters to establish a baseline performance.

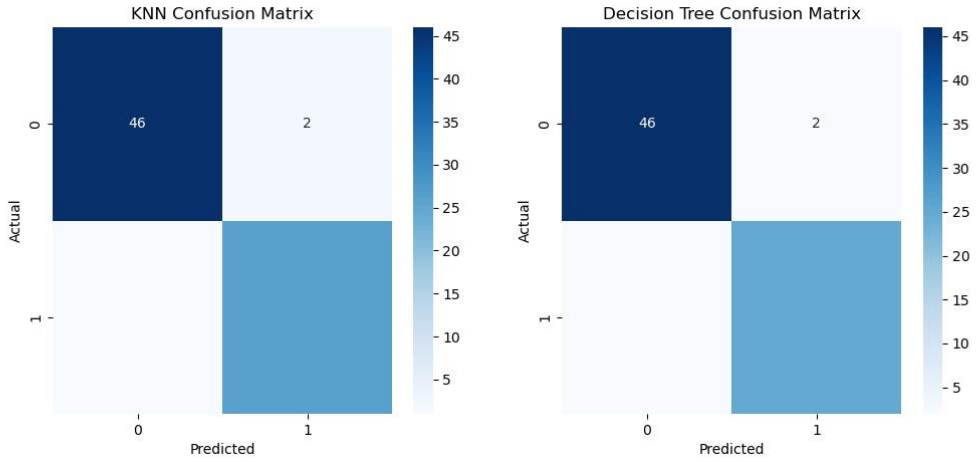
3. **Hyperparameter Tuning:** Model parameters were fine-tuned to improve performance. For the KNN classifier, parameters such as the number of neighbors (**n_neighbors**), distance weighting (**weights**), and distance metric (**p**) were optimized. For the Decision Tree classifier, parameters such as tree depth and splitting criteria were adjusted.

4. **Model Validation:** Models were validated using k-fold cross-validation to ensure robustness and prevent overfitting.

Model Performance Metrics:

- **Accuracy:** Proportion of correctly predicted instances.
- **Precision:** Proportion of true positive predictions among all positive predictions.
- **Recall:** Proportion of true positive predictions among all actual positives.
- **F1-Score:** Harmonic mean of precision and recall, providing a balanced measure.
- **Cross-Validation Accuracy:** Average accuracy across multiple folds of cross-validation, ensuring the model's generalizability.

Figure 6: KNN and Decision Tree Confusion Matrix



Results

KNN Classifier

- **Accuracy**: 0.9600
- **Precision**: 0.9607
- **Recall**: 0.9600

- **F1-Score**: 0.9602
- **Cross-Validation Accuracy**: 0.9800

Decision Tree Classifier

- **Accuracy**: 0.9467
- **Precision**: 0.9467
- **Recall**: 0.9467
- **F1-Score**: 0.9467
- **Cross-Validation Accuracy**: 0.9398

Discussion

The evaluation metrics indicate that the KNN classifier outperforms the Decision Tree classifier in predicting user interest in sports. The KNN model achieved higher accuracy, precision, recall, and F1-score compared to the Decision Tree model. Additionally, the cross-validation accuracy of the KNN model was significantly higher, demonstrating its robustness and reliability.

From the data exploration, it was observed that:

- A strong positive correlation between 'Nature' and 'Picnic' reviews suggested that users interested in nature-related activities are also likely to engage in picnics.
- A negative correlation between 'Religious' and 'Theatre' reviews indicated differing user preferences for these categories.

The feature selection process further validated the effectiveness of the KNN classifier. The selected features for both models were similar, indicating that the 'Picnic', 'Theatre', 'Shopping', and 'Nature' categories are most relevant for predicting sports interest.

Conclusion

In conclusion, the KNN classifier is recommended for predicting user interest in sports based on user reviews in various categories. Its superior performance across multiple evaluation metrics and robustness in cross-validation make it the preferred model for this task. Future work may involve exploring additional classification algorithms and feature engineering techniques to further improve model performance.

References

- UCI Machine Learning Repository. (n.d.). BuddyMove Data Set. Retrieved from https://archive.ics.uci.edu/dataset/476/buddymove+data+set
- Scikit-learn documentation. (n.d.). Retrieved from https://scikit-learn.org/stable/documentation.html
- HolidayIQ.com. (n.d.). User Reviews on Points of Interest in South India. Retrieved from https://www.holidayiq.com