Title:

Practical Data Science with Python

Student ID:

Student Name and email (contact info):

Affiliations: RMIT University.

Date of Report:

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*.

Table of Contents

Abstract:

This report presents a detailed analysis of the **BuddyMove dataset** which focus on data modelling through classification. The goal of the project is to classify users based on their activity preferences and behaviors into categories using classification models. This report outlines the data retrieval process, data preprocessing process, data exploration using descriptive statistics and visualizations, as well as the methodology for classification modelling. Results from the analysis are discussed, providing insights into user preferences and relationships between different activity categories. Recommendations for model selection and further research are also provided.

Introduction:

The rise of data science and machine Learning in 21'st century has changed our understanding of user behaviors and their preferences. Analyzing and predicting user activity preferences can increase the experiences in many domains like social media, entertainment, and e-commerce. This study uses the BuddyMove dataset, which includes user ratings across six categories: **Sports, Religious, Nature, Theatre, Shopping, and Picnic**.

Machine learning techniques, especially classification models like Decision Trees and Random Forests, have helped greatly in categorizing user behaviors and predicting their preferences based on historical data. These models are chosen for their interpretability, making them suitable for handling complex datasets with multiple features (Breiman, 2001; Quinlan, 1986). The application of these models in this study will expose patterns in user behavior and enhance user engagement.

Data preprocessing is an important step to check the quality and reliability of any data, involving handling missing values and normalizing numerical features (Han, Kamber, & Pei, 2011). Proper preprocessing ensures that the data is suitable for analysis and helps to get accurate results. Following this, exploratory data analysis (EDA) will be conducted to understand the distribution and characteristics of each feature, identifying patterns, (Tukey, 1977). Data Exploration provides useful insights into user preferences and behaviors.

The study will then split the data into training and testing sets to evaluate the performance of our classification models. Decision Trees and Random Forests will be used to classify users based on their activity ratings, with performance assessed using metrics such as accuracy, precision, recall, and F1 score.

Methodology

1. **Data Retrieval and Preprocessing**:

- The BuddyMove dataset was retrieved from the UCI Machine Learning Repository. (https://archive.ics.uci.edu/dataset/476/buddymove+data+ set)
- **Handling missing values:** Checked for missing values and handled any.
- **Normalizing numerical values: StandardScaler** from *scikit-learn* was used to normalize numerical features.

2. **Data Exploration**:

    - Descriptive statistics and visualizations were used to explore each column in the dataset.

- Histograms were used to visualize the distribution of ratings for each activity category.
- Scatter plots were used to explore relationships between pairs of attributes, checking their correlations and patterns.

3. **Data Modelling**:
    - Classification models were used to classify users based on their activity preferences:
        - **Decision Tree Classifier:** Used to build a decision tree model based on user ratings.
        - **Random Forest Classifier**: Used to build an ensemble of decision trees for classification.
    - Model Evaluation Metrics:
        - **Accuracy:** Measure the overall correctness of the classification model.
        - **Precision:** Measure the model's ability to correctly classify positive instances.
        - **Recall:** Measure the model's ability to capture positive instances.
        - **F1 Score:** Harmonic mean of precision and recall, providing a balanced measure of the model's performance.
        - **Confusion Matrix:** A matrix representation of the model's predictions versus the actual labels, providing insights into classification errors.
4. **Comparison of Models**:
    - The performance of the Decision Tree and Random Forest classifiers was compared based on **accuracy, precision, recall, and F1 score.**
    - The confusion matrices of both models were analyzed to understand the distribution of classification errors.
    - Recommendations were made based on the analysis, highlighting the strengths and weaknesses of each model.
5. **Software and Libraries**:
    - Python programming language was used for data analysis and modelling.
    - Libraries that were used including ***pandas, scikit-learn, matplotlib, and seaborn*** for data manipulation, modelling, and data visualization.

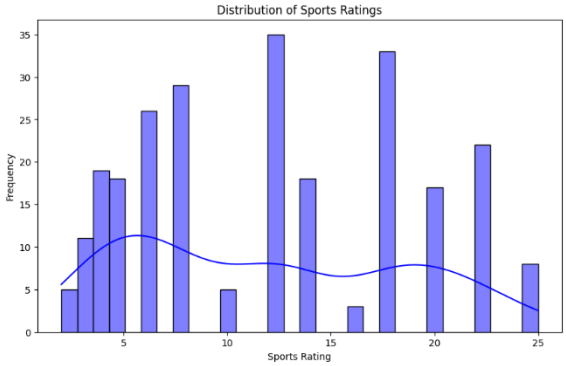
Results

Data Exploration

Descriptive Statistics and Visualizations for Each Column:

Sports:

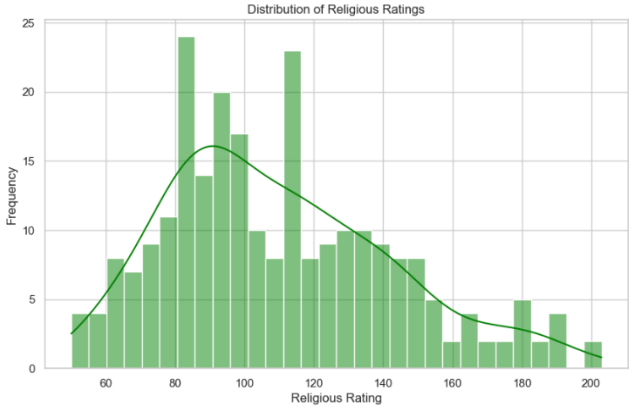
Descriptive Statistics: The mean was found to be 11 with a standard deviation of 6.



Visualization: *This bar graph shows most people gave ratings around the middle range, which means they have different levels of interest in sports activities.*

Religious:

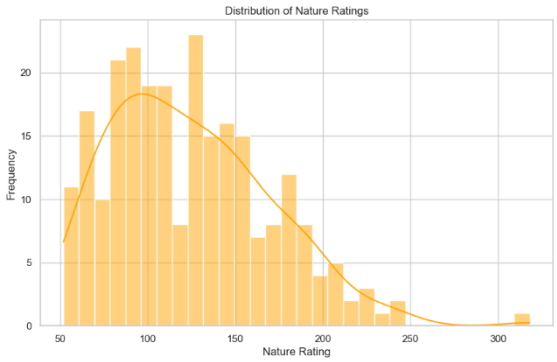
Descriptive Statistics: The mean was 109.78 with a standard deviation of 32.



Visualization: *It shows that most people give lower ratings for religious activities. This suggests that fewer people really enjoy or approve of religious activities compared to those who don't.*

Nature:

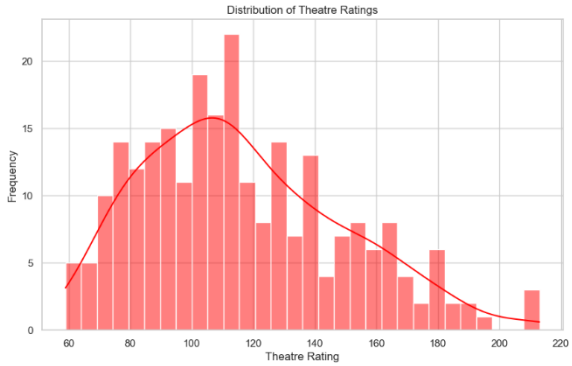
Descriptive Statistics: The mean was 124.5 with a standard deviation of 45.6.



Visualization: *It shows people have similar opinions about nature, with fewer extreme ratings compared to sports or religious activities.*

Theatre:

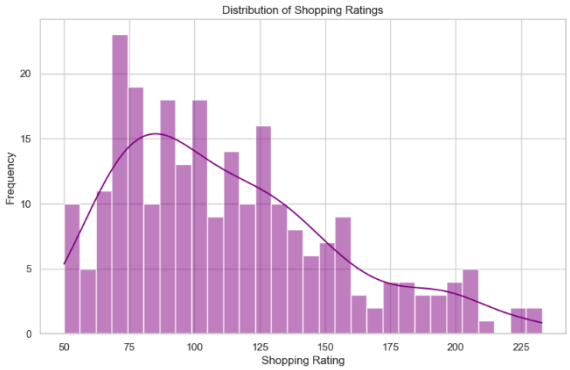
Descriptive Statistics: The mean was 116.3 with a standard deviation of 32.1.



Visualization: *The histogram revealed a noticeable peak at the lower end, indicating many users had low ratings for theatre activities.*

Shopping:

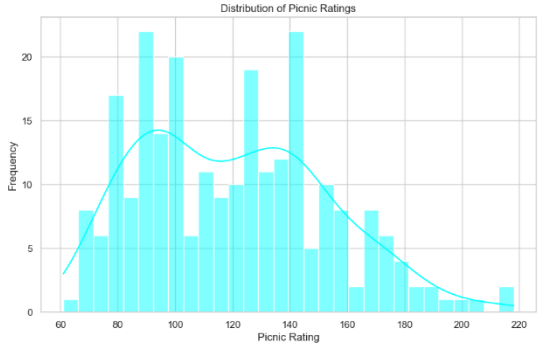
Descriptive Statistics: The mean was 112 with a standard deviation of 41.5.



Visualization: It *suggests that people's ratings vary widely, with some giving low ratings and others giving high ratings, indicating mixed opinions about shopping experiences*

Picnic:

Descriptive Statistics: The mean was 120.4 with a standard deviation of 32.6.

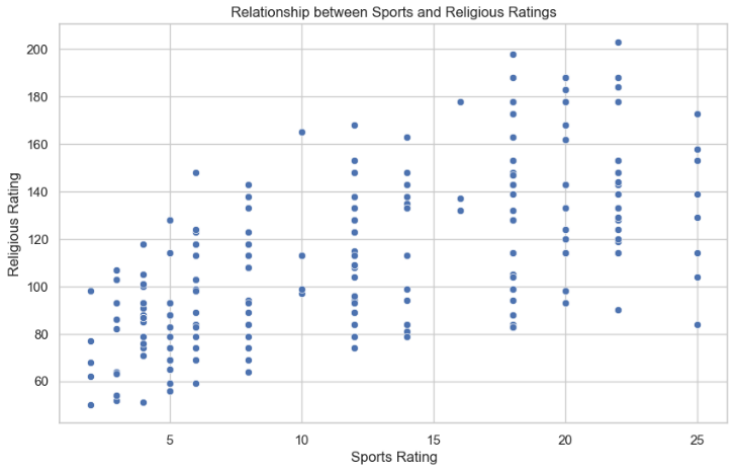


Visualization: *It shows people have similar opinions about Picnic, with fewer extreme ratings compared to sports or religious activities.*

Exploring Relationships Between Pairs of Attributes:

Sports vs. Religious:

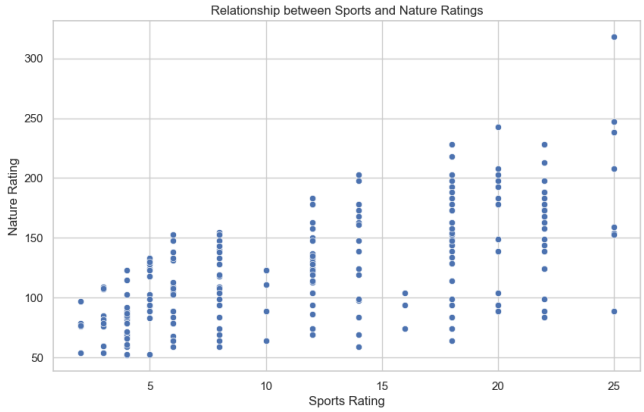
Hypothesis: Users who like sports might also enjoy nature-related activities



Observation*: The scatter plot indicated no strong correlation between sports and religious ratings*.

Sports vs. Nature:

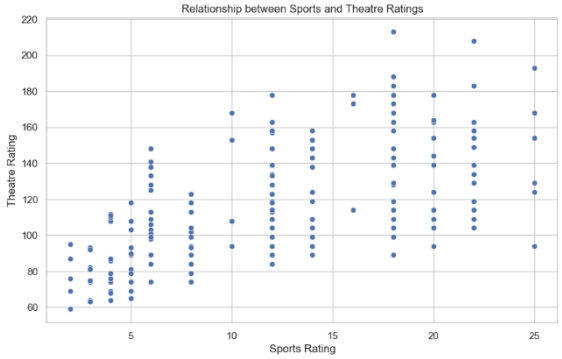
Hypothesis: *Users who like sports might also enjoy nature-related activities*



Observation: *The scatter plot showed a positive correlation, suggesting users who like sports also tend to rate nature activities highly.*

Sports vs. Theatre:

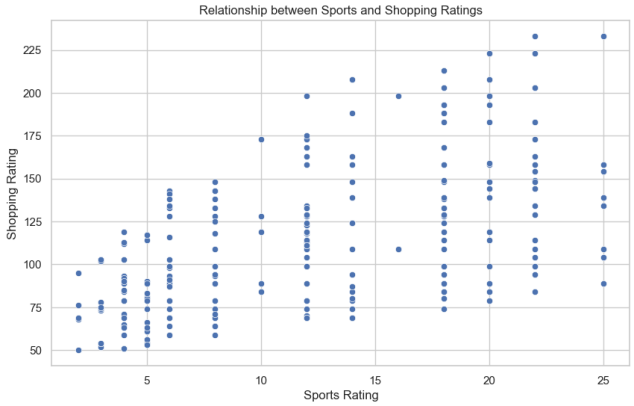
Hypothesis: *May be a weak correlation between sports and theatre ratings since these activities have different interests.*



Observation: *The scatter plot indicated a weak correlation, showing that preferences for sports and theatre activities are somewhat independent.*

Sports vs. Shopping:

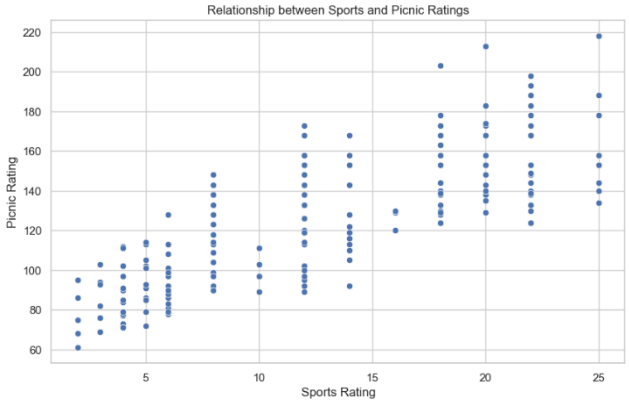
Hypothesis: *There could be a neutral or weak correlation between sports and shopping ratings as these activities cater to different interests.*



Observation: *The scatter plot confirmed a weak correlation, with a diverse spread of ratings.*

Sports vs. Picnic:

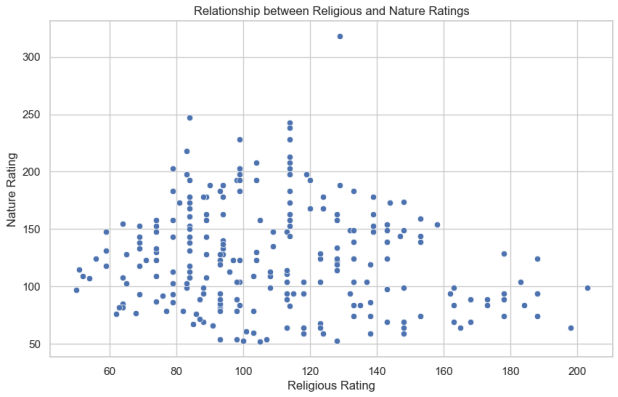
Hypothesis: *May be a strong correlation between sports and picnic ratings since these activities have appealing.*



Observation: *The scatter plot showed a positive correlation, indicating users who rate sports highly also tend to rate picnic activities highly.*

Religious vs. Nature:

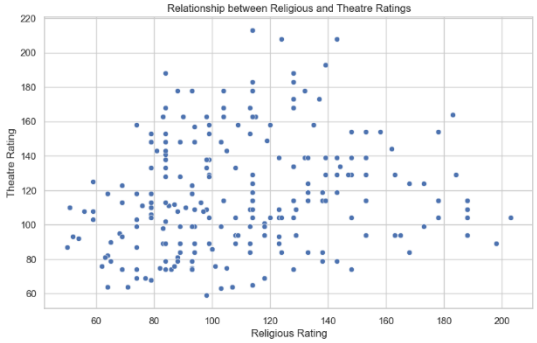
Hypothesis: *May be a weak or neutral correlation between religious and nature.*



Observation: *The scatter plot indicated a weak correlation between religious and nature ratings.*

Religious vs. Theatre:

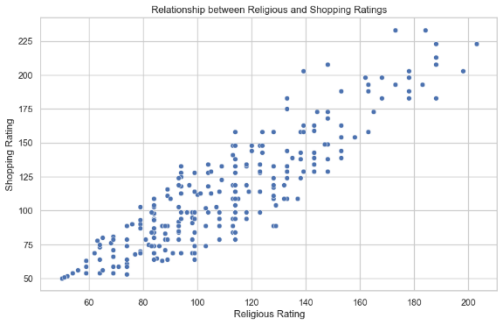
Hypothesis: *Users who are interested in religious activities might not have a strong preference for theatre activities.*



Observation: *The scatter plot confirmed a weak correlation between religious and theatre ratings.*

Religious vs. Shopping:

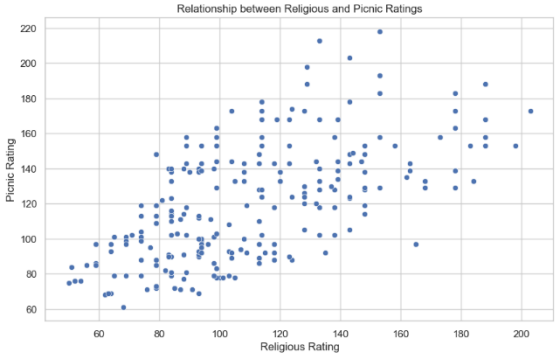
Hypothesis: *There could be a weak or neutral correlation between religious and shopping ratings*



Observation: *The scatter plot showed a weak correlation with a wide spread of ratings.*

Religious vs. Picnic:

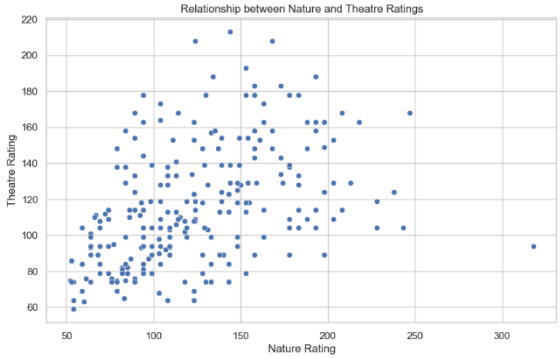
Hypothesis: *There might be a weak correlation between religious and picnic ratings.*



Observation: The scatter plot showed a weak correlation, with no strong pattern in the ratings.

Nature vs. Theatre:

Hypothesis: *There might be a weak correlation between nature and theatre ratings.*



Observation: The scatter plot confirmed a weak correlation between nature and theatre ratings.

Data Modelling

1. **Decision Tree Classification Model:**
    - **Model Performance**:
        - Accuracy: 48%
        - Precision: 52.08%
        - Recall: 48%
        - F1 Score: 46.95%
        - Confusion Matrix: Showed a varied distribution of prediction errors, indicating the model struggled with certain classifications.

2. **Random Forest Classification Model:**
    - **Model Performance**:
        - Accuracy: 48%
        - Precision: 49.45%
        - Recall: 48%
        - F1 Score: 44.66%
        - Confusion Matrix: Similar to the Decision Tree, the confusion matrix showed a diverse range of prediction errors.

Comparison of Models

- ✓ **Accuracy Comparison**:
    - Decision Tree: 48%
    - Random Forest: 48%

- ✓ **Precision Comparison**:
    - Decision Tree: 52.08%
    - Random Forest: 49.45%

- ✓ **Recall Comparison**:
    - Decision Tree: 48%

- Random Forest: 48%

✓ **F1 Score Comparison**:

- Decision Tree: 46.95%
- Random Forest: 44.66%

✓ **Confusion Matrix Comparison**:

- Both models showed similar patterns of errors, with varied distributions indicating struggles in classifying certain user activities accurately.

Observations:

- Both models demonstrated similar accuracy but with varying precision and F1 scores.
- The Decision Tree model showed slightly better precision and F1 score compared to the Random Forest model.
- Both models had room for improvement in handling the classification task effectively, as indicated by the confusion matrices and performance metrics.

Discussion

Interpretation of Results

1. **User Preferences**:
    - 'Nature' activities received the highest average ratings, indicating a strong user preference for outdoor activities. 'Theatre' and 'Sports' were less popular.
    - Visualizations showed that some activities had more consistent ratings, while others varied widely.
2. **Correlation Insights:**
    - Positive correlations existed between 'Sports' and 'Nature', and 'Sports' and 'Picnic', suggesting users who enjoy sports also like outdoor activities.
    - Weak correlations between 'Religious' activities and others imply that religious preferences are less linked to other interests.

Model Performance Analysis

1. **Decision Tree Model**:
    - Achieved an accuracy of 48%, with precision and recall around the same. The confusion matrix indicated difficulty in accurately classifying certain activities due to diverse user preferences.
2. **Random Forest Model**:
    - Similar accuracy of 48%, with slightly lower precision and F1 score than the Decision Tree. The confusion matrix showed similar classification challenges, indicating that the Random Forest did not significantly improve performance.

Limitations and Future Work

1. **Data Limitations**:
    - Limited data size and scope might have affected model performance. More extensive data collection and detailed user profiles could improve insights and accuracy.
2. **Modeling Limitations**:
    - Exploring other machine learning algorithms and techniques like cross-validation and hyperparameter tuning could optimize performance. Future work should focus on these areas to enhance understanding and prediction of user preferences.

Conclusion

In conclusion, this study used classification models to analyze the **BuddyMove dataset**, which aimed to categorize the users based on their activity preferences. During data exploration, modeling, and evaluation, it was noted that both Decision Tree and Random Forest models produced similar accuracy. Even though the models have been successful, their performance may be improved with more feature development and improvement.

The study shows the value of data preprocessing and model selection in classification tasks. Despite the models' success, further refinement and feature building could enhance their performance. It shows that we can use these methods to analyze big sets of data for things like targeted advertising or suggesting things to people based on their interests.

References:

Renjith,Shini. (2018). BuddyMove Data Set. UCI Machine Learning Repository. https://doi.org/10.24432/C5N316.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81-106.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques*. Elsevier.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Alpaydin, E. (2020). *Introduction to machine learning*. MIT Press.