AIC-501 Supervised and Unsupervised Machine Learning
Assessment Activity AIC-501:00010
Supervised Learning for Structured Data

Name: (Win) Thanawin Pattanaphol Student ID: 01324096

1)
   a) The specific features would be the following:
      1) Demographic Features
         A) Age
         B) Gender
         C) Nationality
         D) Income Level
      2) Travel Behavior Features
         A) Type of trip
         B) Length of stay in Thailand
         C) Method of transsportation
      3) Preference Features
         A) Interests (history, food, nature)
         B) Activities (hiking, diving, sightseeing)
         C) Favorite type of accommodations (luxury, budget)
      4) Behavior Features
         A) Time spent on social media
         B) Number of reviews or ratings left for destinations asnd activities,.
   b) There are several ways that we can collect these information from the tourists that are coming to visit our country.
      1) Travel Forums
      2) Online Polls
      3) On-site forums or forms

2) In categorizing these data, there are several steps that we can ensure the accuracy of such data
   a) Clearly drawn labels – What categories are we using the label our data?
   b) Define data quality standards – What data is considered valid and not valid?
   c) Validate data that have been received as some data might not be correct such as wrong data type or information.
   d) Clean up data regularly – To clear out any outliers or incorrect information.

3) Data requirement estimation with an accuracy level of at least 80% would at at a minimum of 10,000 – 20,000 tourist records as the majority of data surveys in Thailand are done at this rate.

4) Attached in Canvas

5) I would use Random Forests as it is a suitable model for handling both categorical and continuous data in classification tasks. It performs well on complex datasets and doesn't require heavy preprocessing as well as handles non-linear relationships effectively, and works efficiently even with minimal feature engineering

AIC-501 Supervised and Unsupervised Machine Learning
Assessment Activity AIC-501:00010
Supervised Learning for Structured Data

Name: (Win) Thanawin Pattanaphol Student ID: 01324096

6) In terms of data splitting, I will be using the 80/20 split where 80% of the data is used for training and 20% for testing. This is a mainstream method that works well for most classification problems, providing sufficient data for training while minimizing the risk of overfitting.

7) Data biases may come from some groups, such as certain nationalities, income levels, or age brackets, might be overrepresented or underrepresented in the dataset.

8)

Class	Precision	Recall
Adventure	0.71	0.81
Cultural	0.70	0.70
Relaxation	0.86	0.90
Shopping	0.95	0.95
Eco-tourism	0.89	0.85

- Shopping shows the highest precision and recall, meaning this category is easiest to distinguish with minimal misclassification.
- Cultural exhibits the lowest precision and recall, suggesting significant overlap with other categories, particularly Adventure and Relaxation.
- Adventure and **Eco-tourism** show moderate precision and recall, indicating some overlap but reasonably accurate classifications.
- Relaxation performs excellently, with high precision and recall, indicating that it is well-differentiated from other categories.

9)
a) **Highest False Positive Rate**: Adventure with 8.64% of instances misclassified as another category.
b) **Highest False Negative Rate**: Cultural with a 30% misclassification rate (i.e., instances where cultural travelers were categorized as other types).

10) There are several ways to further improve
a) Distinctive Feature Addition: Introduce additional features that can better differentiate overlapping categories. For example, adding "physical activity level" could improve the distinction between Adventure and Relaxation, and introducing "historical site preference" could better differentiate cultural tourists.
b) Data Augmentation: This is for categories with lower accuracy or higher misclassification rates. Adjusting class weights in the model can help represent more categories, improving overall recall and precision balance.