**Data and Data Preprocessing**

**Problem 1: Types of Attributes (14 points)**

Classify the following attributes as nominal, ordinal, interval, ratio. **Explain why.**

1. Rating of an Amazon product by a person on a scale of 1 to 5

We utilize **ordinal data types** to establish ranking and scaling metrics for Amazon products. Ordinal data allows us to assign a meaningful order to products, enabling effective performance evaluation and attribute scaling. This structured approach enhances our analysis in the dynamic Amazon marketplace.

1. The Internet Speed

Internet speed is an changing value and needs to be in **ratio data type**, as it can have a true zero value and can be convert to different metric it can be megabytes per seconds to kilobytes per second

1. Number of customers in a store.

**Interval data types** because they accommodate random numbers that can change at any given moment. With interval data, we can effectively represent and work with values that vary unpredictably, making it a suitable choice for our needs. This flexibility allows us to handle data that may fluctuate over time without the constraints associated with other data types.

1. UCF Student ID

We use **nominal datatype** as they need to be unique and catered to each student who might be in the university.

(e) Distance

We employ **ratio data types** for distance measuring metrics because they inherently maintain a true zero value and allow for meaningful comparisons and mathematical operations between different metrics.

1. Letter grade (A, B, C, D)

They need to be in the **ordinal data types** as the letter grade need to be in a increasing manner of importance, so we need an ordinal data types system where the grades are in an increasing manner

1. The temperature at Orlando - is suitable for representing the temperature in Orlando. This choice aligns with the nature of temperature data, which can fluctuate within specific ranges and is subject to continuous variations. By using an interval data type, we effectively capture the variability and random fluctuations in temperature, allowing us to account for the ever-changing nature of this environmental data. This approach provides a more accurate representation of Orlando's temperature conditions.

**Problem 2: Exploring Data Preprocessing Techniques (26 points)**

Read the solution post of the Kaggle Titanic Dataset:

https://www.kaggle.com/code/preejababu/titanic-data-science-solutions. Run the code and reproduce the data preprocessing and classification modeling steps.

Q1 (Reproduce): Please read, understand, run the code and reproduce the model accuracies.

Please briefly explain whether you can reproduce the classification accuracies of 'Support

Vector Machines', 'KNN', 'Logistic Regression', 'Random Forest', 'Naive Bayes', 'Perceptron',

'Stochastic Gradient Decent', 'Linear SVC', 'Decision Tree'. (10 points)

[Link](https://github.com/Sparsh-Palkhiwala/ASU/blob/main/CSE%20572%20-%20DM/Hw_1/titanic-data-science-solutions.ipynb)

Q2 (Improve): Is the data preprocessing process proposed in the Kaggle post the best preprocessing solution? If yes, please explain why. If not, can you leverage what you learned in the class and your previous experiences to improve data processing, to obtain better accuracies for all these classification models? Describe what is your improved data preprocessing, and what are your improved accuracies? (16 points)

[Link](https://github.com/Sparsh-Palkhiwala/ASU/blob/main/CSE%20572%20-%20DM/Hw_1/titanic-dataset-preprocessing%20(1).ipynb)

* Feature Engineering:
  + Extracted the 'Deck' from the 'Cabin' feature and filled missing values with 'Unknown'.
  + Extracted titles from passenger names and categorized them.
  + Created a 'Family Size' feature by combining 'Parch' and 'SibSp'.
* Data Transformation:
  + Standardized numerical features ('Age' and 'Fare').
  + Encoded categorical features ('Sex', 'Embarked', 'Deck', 'Title').
* One-Hot Encoding:
  + Applied one-hot encoding to categorical features ('Embarked', 'Pclass', 'Title', 'Deck').
* Data Alignment:
  + Ensured consistent columns in both training and test sets.
* Model Building and Evaluation:
  + Trained and evaluated various models (e.g., Logistic Regression, Random Forest).
  + Identified Random Forest and Decision Tree as top-performing models based on accuracy.
* Test Set Prediction and Submission:
  + Used the Random Forest model to predict survival on the test set.
  + Created a submission file for Kaggle.

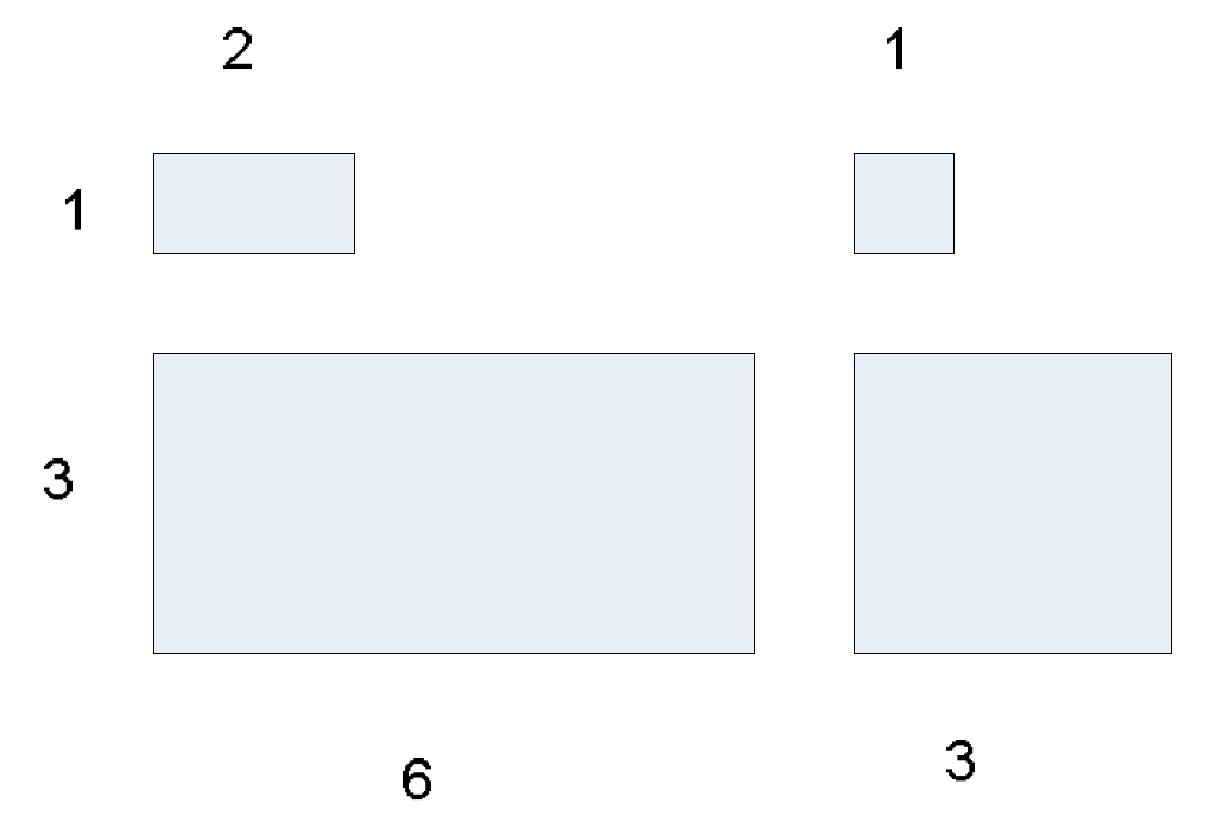
We also tried tuning the hyper parameters of the model using GridSearchCV but unfortunately were not able to find improve any form of accuracy or measuring metrics.

A screenshot of a computer

Description automatically generated

**Problem 3: Distance/Similarity Measures (10 points)**

Given the four boxes shown in the following figure, answer the following questions. In the diagram, numbers indicate the lengths and widths and you can consider each box to be a vector of two real numbers, length and width. For example, the top left box would be (2,1), while the bottom right box would be (3,3). Restrict your choices of similarity/distance measure to Euclidean distance and correlation. **Please explain your choice.**



Which proximity measure would you use to group the boxes based on their shapes (length-width ratio)?

To group the boxes based on their shapes, particularly the length-to-width ratio, a suitable proximity measure would be the **correlation measure**.

We use a correlation measure as we want to determine the proximity measure , as the ratio of the length of box and width of box is same , we should prefer to use a correlation measure. This ratio is consistent across the boxes and thus correlation measure can effectively capture this similarity

Which proximity measure would you use to group the boxes based on their size?

We prefer to use an **Euclidean measure** to group the boxes according to their size. This Euclidean measure is commonly used to quantify the separation between data points and an Euclidean measure can effectively capture that

**Please submit a PDF report. In your report, please answer each question with your explanations, plots, results in brief. DO NOT paste your code or snapshot into the PDF. At the end of your PDF, please include a website address (e.g., Github, Dropbox, OneDrive, GoogleDrive) that can allow the TA to read your code if any.**