# CSE 572: Data Mining
# Homework 1
# Data and Data Preprocessing

**Name:** Dev A Patel **ASU ID:** 1229417087

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

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

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

The rating of an Amazon product is an **ordinal** attribute as the rating can be ranked in an order, high to low or vice-versa.

(b) The Internet Speed

It is a **ratio** attribute. Internet speed is ratio between Megabits transmitted to time taken (seconds) - Mbps.

(c) Number of customers in a store.

It's a **ratio** attribute as it follows the conditions for a ratio. We can perform all mathematical operations. It has equal inteval, for example the difference between 20 and 40 customers is same as the difference between 300 and 320 customers.

(d) UCF Student ID

It's a **nomial** attribute as it serves the purpose of being an unique identifier and has no quantitative value.

(e) Distance

It is **ratio** attribute as it has consistent intervals, and we can perform algebric operations on it. For example, 20 miles is twice 10 miles.

(f) Letter grade (A, B, C, D)

It is **ordinal** attribute. Letter grades can be ordered like A>B>C>D.

(g) The temperature at Orlando

It is an **interval** attribute. It has equal intervals, possess an arbitary zero point - temperature can truly be zero (hence not ratio), allows quantitative comparision, and we can perform mathematical operations. Also there are not ratios, as we can’t multiply or divide them meaningfully.

## 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)**

Given the titanic dataset, we need to predict whether the person survived the crash based on the attributes like 'PassengerId' 'Survived' 'Pclass' 'Name' 'Sex' 'Age' 'SibSp' 'Parch' 'Ticket' 'Fare' 'Cabin' 'Embarked'. The type of each attribute has been identified. Features containing NULL attributes and distribution of numerical and categorical attributes were identified. They dropped certain columns such as ticket, cabin, name, and passengerid as they will not logically contribute for final prediction. New features were also created. Then all mentioned algorithms were applied.

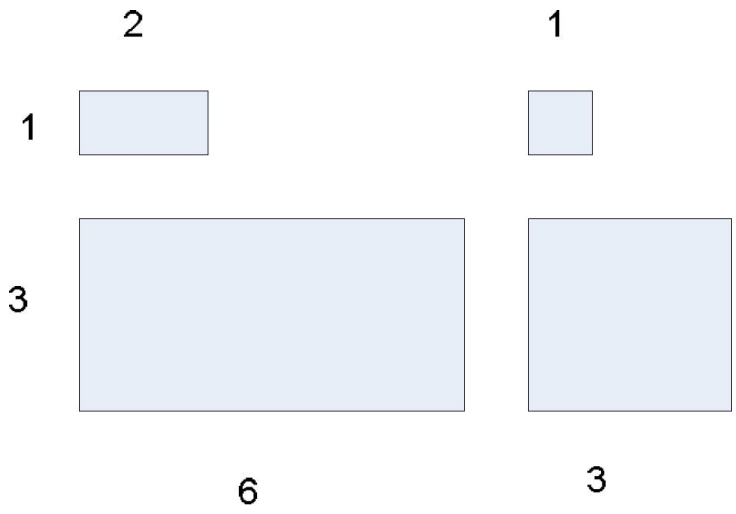
I was able to reproduce same results when I ran the code on my system, fetching all accuracies.

**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)**

The proposed preprocessing techniques in the kaggle code seems to be most suitable as the analysis has been done considering all the aspects. Creating and dropping new features, analysing correlation among numerical and ordinal features, and converting features to required datatype for prediction seems to be done in a better way as compared to other kaggle codes.

## 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)?**

In order to group the boxes based on shapes we can use **correlation.**

$$corr(x,y) = covariance(x,y)/(sd(x).sd(y))$$

If the value of correlation is closer to 1, than it means that the boxes are closely related, and if correlation is closer to 0 it means that the boxes are less related. Correlation measure linear relationship between 2 variables, here it is length and width. If correlation is close to 1 it will mean boxes have similar L to W ratio and hence similar shapes to be grouped.

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

In order to group boxes based on size, Euclidian distance can be used. It will give us the spatial distance between the boxes, comparing the magnitude of the dimensions. If the distance is low, it implies that boxes are close and we can group them.

**Github Link:** https://github.com/dpate243/CSE572-Data-Mining