**Homework Assignment #1**

**Due Date: 02.28.25 – Friday via email**

**Problem 1.**

Discuss whether or not each of the following activities is a data mining task.

(a) Dividing the customers of a company according to their gender.

(b) Dividing the customers of a company according to their profitability.

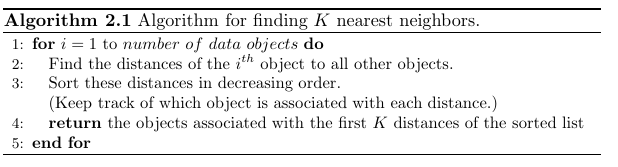
(c) Computing the total sales of a company. No. Again, this is simple accounting.

(d) Sorting a student database based on student identification numbers.

(e) Predicting the outcomes of tossing a (fair) pair of dice.

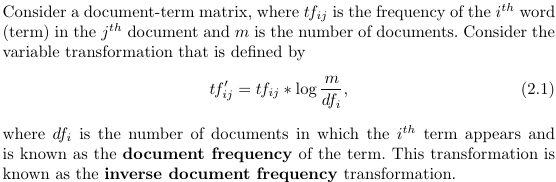
(f) Predicting the future stock price of a company using historical records.

**Problem 2.**



1. Describe the potential problems with this algorithm if there are du plicate objects in the data set. Assume the distance function will only return a distance of 0 for objects that are the same.
2. How would you fix this problem?

**Problem 3.**



(a) What is the effect of this transformation if a term occurs in one document?

(b) What might be the purpose of this transformation?

**Problem 4.**

Proximity is typically defined between a pair of objects.

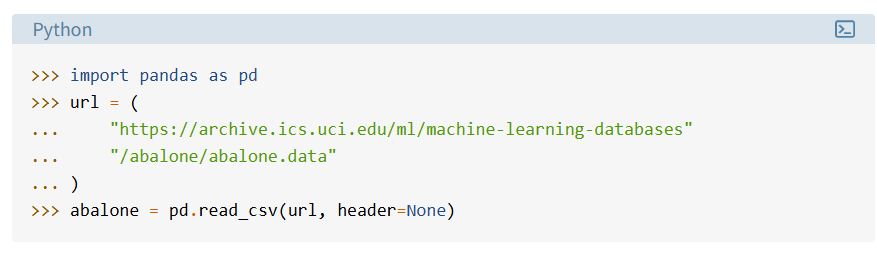
1. Define two waysin which you might define the proximity among a group of objects.
2. How might you define the distance between two sets of points in Euclidean space?
3. How might you define the proximity between two sets of data objects? (Make no assumption about the data objects, except that a proximity measure is defined between any pair of objects.)

**Problem 5**

The goal of the project is to develop a model that can predict the age of an abalone based purely on the other physical measurements. This would allow researchers to estimate the abalone’s age without having to cut its shell and count the rings.

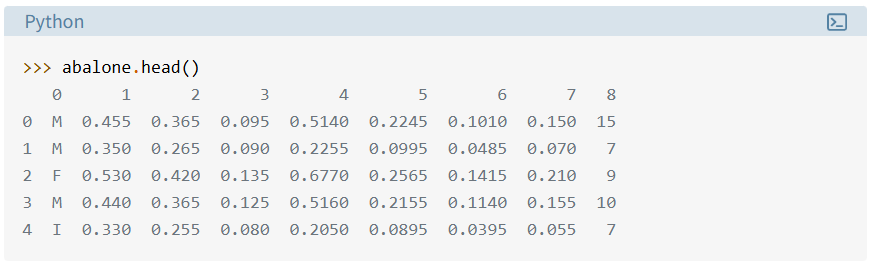
Hint: You’ll be applying a kNN to find the closest prediction score possible.

1. Work with the [Abalone Dataset](https://archive.ics.uci.edu/ml/datasets/abalone) (<https://archive.ics.uci.edu/dataset/1/abalone>)
2. Download it and use [pandas](https://realpython.com/learning-paths/pandas-data-science/) to import the data into Python, but it’s even faster to let pandas import the data directly for you.
3. It is recommended to install Python with Anaconda. The Anaconda distribution comes with many important packages for data science. For more help with setting up your environment, you can check out  [Setting Up Python for Machine Learning on Windows](https://realpython.com/python-windows-machine-learning-setup/).
4. You can import the data using pandas as follows:



In this code, you first [import](https://realpython.com/python-import/) pandas, then you use it to read the data. You specify the path to be a URL so the file will be fetched directly over the Internet.

To make sure that you’ve imported the data correctly, you can do a quick check as follows:



This should show you the first five lines of the Abalone Dataset, imported in Python as a pandas [DataFrame](https://realpython.com/pandas-dataframe/). You can see that the column names are still missing. You can find those names in the [abalone.names](https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/) file on the UCI machine learning repository. You can add them to your DataFrame as follows:

