Foundation of Data Mining

Assignment #1:

- **What is the difference between Data Mining and Statistics?**

  Statistics tests hypothesis, whereas Data Mining searches through all possible hypothesis.

- **What are the differences/similarities between classification and numerical prediction methods?**
  - Differences - Classification is a set of discrete values(i.e. blue, red, etc) whereas numerical is a real value(i.e. set of Real such that domain(0,10) and range is(-100, 100).
  - Similarities - Both are forms of supervised learning

- **What is the difference between a supervised and unsupervised learning? Provide an example of each.**
  - Supervised - Output datasets are provided and the machine learns(or is trained) to produce the given output. This type of learning has feedback, because the given output is given. Example: Digit Recognition

  - Unsupervised(or clustering) - No datasets are provided. The data is clustered into different data sets. No feedback is given. Example: Cyber-security. Detect profiles of what an attack on network would look like.

- **Why might a pruned decision tree that does not fit the data so well be better than an un-pruned one?**
  - Pruning leads to less complexity while also discarding portions of tree that does not lead to better classification because that portion of tree has so few entries.

- **What is the difference between a training set, validation set and a test set?**
  - Training Set - Used to adjust the weights of a neural network. Used in supervised learning.

- Validation Set - Used to signal no more training of machine is needed.
- Test Set - Used for testing the final solution. Buttresses the conclusion that the machine has good predictive power.

- **What is the difference between an Attribute and an Instance?**
    - Attribute - An attribute is a property of an instance
    - Instance - An instance is composed of attributes. It is a snapshot in time or space typically.
    - An analogy for this could be

      Instance : Attribute :: Person :: Phenotype(hair color, height, skin color, etc)