ISE 365/465 Mid-term Exam Review

1. **You should understand the SEMMA / CRISP-DM Modeling framework.**
2. **You must know the function and settings that we covered in class of the following IBM SPSS Modeler and Enterprise Miner Nodes and know how to use them to build a Stream/Diagram and interpret the results:**

   <u>IBM SPSS Modeler</u>

   a. Source Nodes
   b. Record Ops Nodes
      i. Select
      ii. Sample
      iii. Sort
      iv. Merge
         1. Different Types of Merging
      v. Append – Haven't covered in detail
      vi. Distinct
      vii. Aggregate
   c. Field Ops Nodes
      i. Auto Data Prep
         1. Settings
         2. Imputation
      ii. Type
         1. Settings
      iii. Filter
      iv. Derive
         1. CLEM Expressions (strings, dates, mean, etc.)
      v. Filler
         1. CLEM Expressions (strings, dates, mean, etc.)
      vi. Reclassify
      vii. Binning
      viii. Partition
      ix. Field Reorder
   d. Graph Nodes
      i. Graphboard
      ii. Plot
      iii. Distribution
      iv. Histogram

v.

SAS Enterprise Miner

e. Sample Nodes
    i. Creating a SAS Data Set and Library
    ii. File Import Node
    iii. Data Partition Node
    iv. Merge Node – We did not cover in detail, but it is similar to IBM SPSS Modeler with less functionality
f. Explore Nodes
    i. Graph Explore Node
    ii. MultiPlot Node
    iii. StatExplore Node
    iv. Variable Selection Node
g. Modify Nodes
    i. Drop Node
    ii. Replacement Node -
    iii. Transform Variables Node
    iv. Principal Components Node
h. Model Nodes
    i. Decision Tree Node
        1. Gini Index for Splits
        2. Gain Ratio for Splits
        3. Pruning/Stopping
        4. Interpretation
        5. When to use different tree algorithms
    ii. Regression Node (linear regression)
        1. Settings
        2. Interpretation of Results
        3. Assumptions of Linear Regression
i. Assess Nodes
    i. Model Comparison Node -
j. Utility Nodes
    i. SAS Code Node
        1. Proc corr
        2. Define new variable

3. You must know the concepts from Chapters 1, 2, and 3, 4.1, 4.2, 8.1, 8.2, 8.5 in the book and all slides and material from class lectures.
    a. General Concepts from Chapter 1

   b. Normalization, Standardization, Correlation, and Chi-square
   c. Classification vs. Prediction
   d. Cumulative % Captured Charts, Lift Charts, Confusion Matrix and associated calculations for a categorical target variable (e.g. sensitivity, specificity, accuracy, precision), ROC chart
   e. Plus topics above number 3 that are in these chapters
4. Decision Tree lecture material
5. You must know all material from in-class labs and homework.
6. Concepts from the "Data Mining Overview" and "Top Ten Data Mining Mistakes" Readings.