ITCS 6265 met for nine class meetings. The purpose of the class was to review the entire CRISP-DM process while applying the process to a real world study. The class studied the Global Shark Attack database and integrated the database with other data sources.

Research Process:

1. Business Research/Understanding: Students studied journal articles, popular press articles and attended a presentation by Dr. Chuck Bangley, shark researcher, ECU.
2. Data Understanding – exploratory data analysis – https://github.com/AKDDResearch/Shark-Attack/blob/master/DataMining/Results/EDA/EDA.docx
   Data sets used:
   a. GSAF (2009 to present, NC and SC data) merged with all dates 5/01 to 9/31; analysis focused on 6/01 to 8/31 due to class imbalance.
      Github Link : https://github.com/AKDDResearch/SharkAttack/tree/master/DataMining/Data/Sharkfile
   b. Turtle data NC, SC (2009 to 2015): raw data has confidential aspects.
      Github Link: https://github.com/AKDDResearch/SharkAttack/tree/master/DataMining/Data/Turtle%20Activity includes the prepared data on turtle false crawls and nesting activity.
   c. Crab Landings data NC (1994 – 2015, used data for 2009 - 2015): raw data has confidential aspects; prepared data included on
      Github Link https://github.com/AKDDResearch/SharkAttack/tree/master/DataMining/Data/Crab%20landings
   d. NOAA weather and water data – raw data source: East Cribbing Station – location of this station is central to NC/SC attack locations.
      Github Link : https://github.com/AKDDResearch/Shark-Attack/tree/master/DataMining/Data/Weatherdata
   e. Moon Phases - calculated
3. Data Preparation –
   Github Link : https://github.com/AKDDResearch/SharkAttack/blob/master/DataMining/Resources/Data%20Source.docx
   Listed below are common steps in the data preparation process. Steps not applied for this research are indicated as future research areas.
   a. Remove unique variables (ID field), special characters (for WEKA – ‘ “ , . & () etc.)
   b. Resolve data imbalance problem: a random method was used to select 1/3 of the records from “attack = no”, means were verified from the sample set to the actual data via t-test statistic (Github Link : https://github.com/AKDDResearch/Shark-Attack/blob/master/DataMining/Data/filebalancedWEKAfinal.csv).
   c. Normalization of numeric variables (min-max score standardization)
   d. Discretization of numeric variables to categorical variables;
      i. 3 bins
      ii. NOTE: Binning is an art, requiring judgement – where can we insert boundaries to maximize differences between attack = Y/N?
         1. Create histograms, look at boundaries to see if improvement can be made between three bins
         2. Example domain knowledge for binning decisions
         3. Adjust Wind to SW for all south westerly winds, moon phase to Full-New Moon for Full Moon and New Moon, Turbidity
      iii. Flag variables – if you want to use regression with categorical variables then coding them into flag variables is important (future research area)

   e. Analysis of correlated predictor variables performed by domain knowledge factors. (future research area)
   f. Missing variables – procedures including imputation and regression (see Modeling part ii below). Crabs – used Estimation to fill in 2016 values. (Turtles – used previous day to replace missing value) (future research area)
   g. Calculate skewness for numeric variables (future research area)
   h. Identifying outliers (future research area)
   i. Removal of duplicate records – function in R used to remove duplicate records.
   j. Integration – all data integrated by date – see WEKA for final file. Files that were integrated had many ways that the date field was stored. Some work was done to insure that the dates were in the same format.

4. Data Files for Analysis:
   a. Main file: Shark Attacks (Y, N). Records from 2008 to present, months May to September, NC and SC.
   b. Shark Attack Y: secondary file used to describe "Shark Attack = Y" conditions
5. Modeling: Appropriate modeling techniques are first selected and then applied with calibration for specific parameters (such as confidence, etc.). Below please find the modeling techniques used for this study.
   Github Link : https://github.com/AKDDResearch/Shark-Attack/tree/master/DataMining/Results
   a. Exploratory Data Analysis and Modeling:
      i. Description – EDA with WEKA was used to explore and learn characteristics of the data, both raw and prepared.
      ii. Estimation and Prediction: – Regression was used to predict crabs based on O2, temp, etc. and the standard error of the estimate was used to determine the quality of the estimation. The Crab data was only for 2015 and there were many missing values.
      iii. Classification
         1. Discretized file with WEKA and predictor variable Attack = Y/N
         2. Normalized file – CART
         3. Neural Network – (future research) must use min/max normalization on all variables (all must be between 0 and 1) and have target variable (supervised)
      iv. Clustering - WEKA
      v. Association - WEKA
      vi. Hypothesis testing – was average higher than normal for 2015? – two sample t-test for diff in means
6. Evaluation
   a. Evaluation for quality and effectiveness
      i. two sample z test for difference in proportions (see if proportions of records with attack = y/n differ between all data and condensed data for class imbalance problem (ok for flag or two values)
      ii. test for homogeneity of proportions
      iii. ANOVA (future research) is mean value of a continuous variable the same across subsets of data?

    b. Does model achieve objectives?
    c. Is there something not accounted for in the business understanding?
    d. Future research: apply model to other areas with area specific data.
        i. DORSAL
        ii. Surfline data

7. Deployment
    a. Report
    b. Integration into app (DORSALl)
8. SOURCES – Github Link : https://github.com/AKDDResearch/Shark-Attack/tree/master/DataMining/Resources