**Project Modeling Strategy**

* Strategy for dealing with NaNs:
  + Test both ways: drop rows (samples) with NaNs, or replace them with average value from that particular category/disease
* Use random forest and selectKbest to identify the 10 proteins from each of them. Then fit models with all proteins, as well as with just the top 10 proteins
  + If there is overlap between the list of most important proteins between random forest and selectKbest, then we will use the overlapping proteins to fit our models
* Compare kNN and logistic regression with all 49 proteins, as well as the most important proteins as decided above