Project Proposal Group 3

Vedashree Bhandare, Amer Rez, Saketh Saxena, Anand Vishwakarma

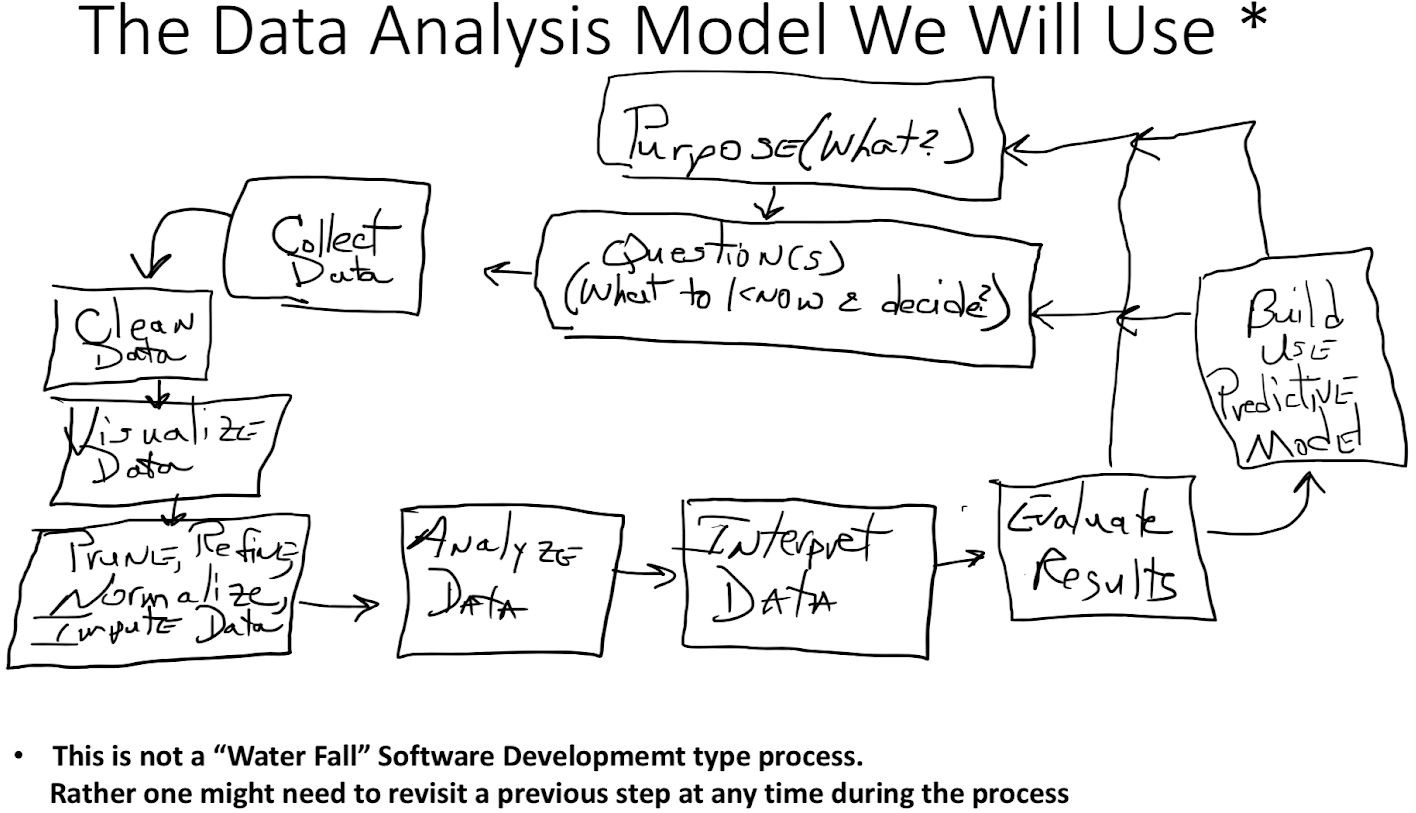
# Project Title

E.) Predict Which Passengers Will Survive The **Titanic using Kaggle** Competition Data.

# Second Choice (plan B):

A.) Predicting Affinity & Activity Relationships Between Small Chemical Molecules and Protein Targets.

We are planning on following the procedure explained in class:



\*Figure taken from Lecture Slides of CS286

INSTRUCTOR COMMENTS:

Decided on Project Title

Predict Which Passengers Will Survive The **Titanic using Kaggle** Competition Data.

Model to use: Suggest using scikit’s SVC (SVM classifier)

CS286 Project Goals:

1. Pick a problem that allows applying at least 80% of the topics & techniques covered in the course to be applied to the problem/task.
   1. For classifier make sure predictors are indep. But Predictors are dep to predicted/target variable (i.e., where the individual survived or not.)
   2. Check that predictors are/can make a significant difference to the dependent/target variable.

Several ways to carry out a & b above. One is to use scatter plots. Another is to use correlation heat maps.

G1: (v1, v2, v3): Check that v1 & v2 are indep. And check that v1 & v3 are indep. and check that v2 & v3 are indep (i.e., check that all combinations of variables to check if they are indep.). This can be accomplished, in part, by scatter plots, correlation heat maps …etc. For example in a scatted plot if you see a linear or non linear relationship between two variables, then they are dep. on each other. Another technique is to generate correlation heat maps, and then you will need to identify some dep. Threshold (e.g., if the correlation is >= 0.2 then you might say the variables are correlated, and if correlation is < 0.2 then they are not. Then you should follow this up with a hypothesis test to determine which variables are or are not dep. on each other. For example, v1-v2: 0.25, v1-v3: 0.1 …… etc. Do a Hyp test to determine which correlation values are significantly different from a 0.2 threshold for a given confidence level (e.g., 95% confidence). 🡨 This is just an example of how one might demonstrate the use of hypothesis testing.

You can end up with groups of variables that are indep. Of each other, but cannot use variables between groups. For example, G2: (v1, v2) are indep. of each other, G3: (v2, v3) are indep. of each other. ~~G4(v1, v3)~~ 🡨 because they might be dep. between each other.

You will need to determine, e.g., for G2, if and the degree to which v1 is dependent on our predicted variable (call it d). Determine the degree of dependence between v1 and d (d = the variable that will indicate survival or non-survival, in other words it is the “class” variable. In binary classifiers, 0 usually represents one class such as non-survival and 1 the other class such as survival) and between v2 and d. If the dependence between v1 and d is non-existent, then there is no sense in including it in the variables that are ultimately used for prediction in your SVC.

Then the idea is to pick a group of indep. variables that each exhibits dep. to the target variable d, and proceed to the next step which can be training and building your SVC. Why we can “the next step can be” is because if the dimension remains too high, then you might need to perform some type of dimension reduction.

You might consider identifying a Beta (or Power) value (i.e., risk of making a Type II error). Then determine how many subjects/samples you need In order to achieve the desired Beta value. OR for the number of samples that you did use, do a reverse calculation to determine what your Beta (or Power = 1 – Beta) value actually is.

1. The Data Analysis steps should be applied to the selected task or problem.
   1. Make sure that you complete, discuss, and cover all of the analysis steps in your report. There should be more than a couple of sentences for each section. Typically I’m looking for a length for each section that is specified in the report template.
2. By early Nov, you should be 30% to 50% complete (e.g., completed data cleaning, visualization, and perhaps determined which predictor variables to use, …etc.). If too far behind that goal, contact the instructor with status/questions.
3. You can use whatever imputation alg that you seem appropriate.