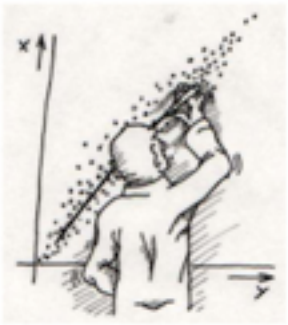
Lab Two: Classification


You are to build upon the predictive analysis (classification) that you already completed in the previous mini-project, adding additional modeling from new classification algorithms as well as more explanations that are inline with the CRISP-DM framework. You should use appropriate cross validation for all of your analysis (explain your chosen method of performance validation *in detail*). Try to use as much testing data as possible *in a realistic manner* (you should define what you think is realistic and why).

This report is worth 20% of the final grade. Please upload a report (one per team) with all code used, visualizations, and text in a single document. The format of the document can be PDF, *.ipynb, or HTML. You can write the report in whatever format you like, but it is easiest to turn in the rendered iPython notebook. The results should be reproducible using your report. Please carefully describe every assumption and every step in your report.

Dataset Selection
Select a dataset identically to the way you selected for the first project work week and mini-project. You are not required to use the same dataset that you used in the past, but you are encouraged. You must identify two tasks from the dataset to regress or classify. That is:

- two classification tasks OR
- two regression tasks OR
- one classification task and one regression task

For example, if your dataset was from the diabetes data you might try to predict two tasks: (1) classifying if a patient will be readmitted within a 30 day period or not, and (2) regressing what the total number of days a patient will spend in the hospital, given their history and specifics of the encounter like tests administered and previous admittance.

Grading Rubric

- Data Preparation (**15 points total**)
    - [**10 points**] Define and prepare your class variables. Use proper variable representations (int, float, one-hot, etc.). Use pre-processing methods (as needed) for dimensionality reduction, scaling, etc. Remove variables that are not needed/useful for the analysis.

    - [**5 points**] Describe the final dataset that is used for classification/regression (include a description of any newly formed variables you created).

- Modeling and Evaluation (**70 points total**)
    - [**10 points**] Choose and explain your evaluation metrics that you will use (i.e., accuracy, precision, recall, F-measure, or any metric we have discussed). Why are the measure(s) appropriate for analyzing the results of your modeling? Give a detailed explanation backing up any assertions.
    - [**10 points**] Choose the method you will use for dividing your data into training and testing splits (*i.e.*, are you using Stratified 10-fold cross validation? Why?). Explain why your chosen method is appropriate or use more than one method as appropriate.
    - [**20 points**] Create three different classification/regression models (e.g., random forest, KNN, and SVM). Two modeling techniques must be new (but the third could be SVM or logistic regression). Adjust parameters as appropriate to increase generalization performance using your chosen metric.
    - [**10 points**] Analyze the results using your chosen method of evaluation. Use visualizations of the results to bolster the analysis. Explain any visuals and analyze why they are interesting to someone that might use this model.
    - [**10 points**] Discuss the advantages of each model for each classification task, if any. If there are not advantages, explain why. Is any model better than another? Is the difference significant with 95% confidence? *Use proper statistical comparison methods*.
    - [**10 points**] Which attributes from your analysis are most important? Use proper methods discussed in class to evaluate the importance of different attributes. Discuss the results and hypothesize about why certain attributes are more important than others for a given classification task.

- Deployment (**5 points total**)
    - [**5 points**] How useful is your model for interested parties (*i.e.*, the companies or organizations that might want to use it for prediction)? How would you measure the model's value if it was used by these parties? How would your deploy your model for interested parties? What other data should be collected? How often would the model need to be updated, etc.?

- Exceptional Work (**10 points total**)
    - You have free reign to provide additional modeling.
    - One idea: grid search parameters in a parallelized fashion and visualize the performances across attributes. Which parameters are most significant for making a good model for each classification algorithm?