1. CREATE MODELS:

Create a logistic regression model and a support vector machine (SVM) model for the classification task involved with your dataset. Assess how well each model performs (use 80/20 training/testing split for your data). Adjust parameters of the models to make them more accurate. If your dataset requires the use of stochastic gradient descent (SGD), then linear kernel only is fine to use. That is, the SGD Classifier is fine to use for optimizing logistic regression and linear support vector machines. For many problems, SGD will be required in order to train the SVM model in a reasonable timeframe.

RESPONSE:

The Chilean earthquake perinatal outcome data set that our group used in Project One was again used here. The data set was reloaded and the cleaning done for Project One (replacement of missing values, renaming of variables, and changing all classification types to int64 or float) was done. Before proceding further, a preliminary logistic model regression was constructed to predict Exposed (whether the newborn infant was exposed *in utero* to the earthquake) from the entire data set. This preliminary model gave an accuracy statistic of 1.000. This indicated over fitting of the data and suggested that some variables were too correlated with each other and/or the outcome. Several modifications of the model parameters were also tried at this point, including changing the value of C from 0.05 up to 50 in increments of 10 and then increments of 5; altering the value of XXX between 12 and 11, increasing the number of iterations up to 20 (because one of the trial conditions indicated it did not reach linearity on at least one iteration). Only changing XXX from 12 to 11 showed any impact. This change increased accuracy from about 0.57 to about 0.60, which is not a very big change. At this point the data were re-examined for correlation using correlation matrix graph plots, correlation heat map plots, and numeric correlation matrix. Several strong correlations were seen. Using a value of the correlation statistic > |0.25|, several variables were dropped (Year, Length, Weight, Trim\_study, Premature. The variables of Day and Month were also dropped as containing irrelevant information. A logistic model was calculated and then an SVM model was prepared. Comparison of the two models showed they had similar accuracies. Logistic Model: accuracy = 0.5xxx, 0.5xxx, and 0.6xxx in the three training runs. SVM accuracy = 0.5XXX. The weights of the SVM model were calculated and plotted. This showed that XX variables had the largest weights. Based on this, the data set was pruned further by removing variables with weights <0.05.. A new logistic regression model and a new SVM model were developed. These models had accuracies very similar to the previous model: Logistic Model: accuracy = 0.5xxx, 0.5xxx, and 0.6xxx in the three training runs. SVM accuracy = 0.5XXX. The weights of the SVM model were calculated and plotted. But this model exhibited large, apparently deleterious decreases in several of what appeared to be major variables (Pondural\_index, apgar\_5, Maternal\_age). Only Head\_circumference was unchanged. These changes led to the conclusion that this model was not as believable as df\_newlite3 as these factors which essentially lost their weight are among those which might be expected to be changed by a stress on pregnancy and therefore contribute to the prediction of exposure to the earthquake. I looked again at df\_newlite4 by including Municipality (a surrogate in the study for socio-economic status) because the earthquake may have affected different municipalities differently. This did not improve the believability of the model, as Municipality took on a weight that virtually negated the others except for Head\_circ, which remained unchanged.

Importantly, none of the above changes in the dataframe contents had any perceivable impact on the accuracy of prediction, which remained around 0.60, or on the confusion matrix. I determined in the end to work with df\_newlite3.

1. MODEL ADVANTAGES:

Discuss the advantages of each model for each classification task. Does one type of model offer superior performance over another in terms of prediction accuracy? In terms of training time or efficiency? Explain in detail.

1. INTERPRET FEATURE IMPORTANCE

Use the weights from logistic regression to interpret the importance of different features for the classification task. Explain your interpretation in detail. Why do you think some variables are more important?.

1. INTERPRET SUPPORT VECTORS

Look at the chosen support vectors for the classification task. Do these provide any insight into the data? Explain. If you used stochastic gradient descent (and therefore did not explicitly solve for support vectors), try subsampling your data to train the SVC model – then analyze the support vectors from the subsampled dataset.