**Human Activity Recognition Project**

* Download the "Human Activity Recognition Using Smartphones Data Set" and its description from [the UCI repository](https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones).
* After you unzip the dataset, read through features\_info.txt. This file describes each feature, its physical significance, and also describes features that are derived from raw data by doing some averaging, or sampling, or some operation that gives a numerical result. What do you notice about the dataset? What kind of guidelines for approaching the dataset do you find?
* In static activities (sit, stand, lie down) motion information will not be very useful.
* In the dynamic activities (3 types of walking) motion will be significant.
* Angle variables will be useful both in differentiating “lie vs stand” and “walk up vs walk down”.
* Acceleration and Jerk variables are important in distinguishing various kinds of motion while filtering out random tremors while static.
* Mag and Angle variables contain the same info as (e.g., are strongly correlated with) XYZ variables. We choose to focus on the latter as they are simpler to reason about. This is a very important point to understand as it results in elimination of a few hundred variables.
* We ignore the band variables as we have no simple way to interpret the meaning and relate them to physical activities.
* mean and std are important, skewness and kurtosis may also so we include all of these.

**Note**: for each of the tasks below, think about how you will fix the data and what the implications / ramifications of your fixes will be. Also, can you think of a way to accomplish several of the cleaning tasks at once?

* Identify and fix the inclusion of ( ) in column names.
* Identify and remove duplicate column names.
* Identify and fix the inclusion of ‘-’ in column names.
* Identify and fix extra ) in some column names.
* Identify and fix inclusion of multiple ‘,’ in column names.
* Identify and fix column names containing “BodyBody”
* Drop 'Body' and 'Mag' from column names.
* Map 'mean' and 'std' to 'Mean' and 'STD'
* Make 'activity' a categorical variable.
* Plot a histogram of Body Acceleration Magnitude (i.e. histogram of all 6 activities) to see how each variable does as a predictor of static versus dynamic activities.
* Split the data into training, test, and validation sets.
* Fit a random forest classifier with 500 estimators to your training set.
* Rank the features by their importance scores. What are the top 10 important features? What is the 10th feature's importance score?
* What is your model's mean accuracy score on the validation and test sets?
* What is your model's precision and recall score on the test set?
* Change the column names to [x0...xn], but keep the last two columns in the dataset named 'subject' and 'activity'.
* Split the data into training, test, and validation sets as follows:
  + Train your model on subjects >= 27
  + Test your model on subjects <= 6
  + Validate your model on subjects >= 21 and < 27
* Fit a random forest classifier with 50 estimators to your training set. What is the oob\_score\_? How does it compare to our previous model's oob score?
* Rank the features by their importance scores. What are the top 10 important features? What is the 10th feature's importance score? How does it compare to the 10th feature's importance score in the last lesson?

Since we are using more features (all of the features, actually) we can expect the importance measure to be spread out more, so individual importance scores are going to be smaller.

* What is your model's mean accuracy score on the validation and test sets?
* Plot your model's confusion matrix.
* What is your model's precision, recall, and F1 score on the test set?

As we (might have) expected, this "black box" random forest performs better than our domain-knowledge driven random forest. But, we have no idea what these top 10 features are.

* See if you can map the top 10 features back to their original column names. (Hint: the number in our new column names should be the column's position.)
* tGravityAcc-mean()-X
* tGravityAcc-mad()-Z
* tGravityAcc-max()-X
* tGravityAcc-max()-Z
* tGravityAcc-min()-X
* tGravityAcc-sma()
* tGravityAcc-energy()-X
* fBodyAcc-bandsEnergy()-25,48.2
* angle(tBodyGyroJerkMean,gravityMean)
* angle(X,gravityMean)