Homework 8

For this homework you will create a python notebook (`.ipynb` file). This file should then be uploaded to wolfware in the assignment link! You must use our jupyter hub for this assignment as we'll be using pyspark.

You are tasked with:

- Finding a data set you can fit supervised learning models with (you cannot use the datasets we've been using in class - there are many places with free data out there such as the UCI machine learning repository and kaggle).
- **Please read your data in via a URL or include the data as a file in your submission.**
- Using a numeric or binary response, fitting three different classes of models and choosing an overall best model.
- Writing a narrative (via a notebook) with explanations and discussions as you go through the above.

Splitting the Data, Metrics, and Models

- Using spark MLlib, split the data into a training and test set.
- Choose and describe a metric you'll be using to judge your models.
- You'll be fitting three different classes of models (of your choice - they can be ones we used in class or ones we didn't cover - see https://spark.apache.org/docs/latest/ml-classification-regression.html for a list of models they have in MLlib). Briefly describe each model (no code or anything here, just concepts and ideas about what the models are doing). These discussions should be clear to someone that knows statistics but doesn't know the modeling type/algorithm!

Model Fitting

Next, you should use Spark MLlib to fit your three different classes models to the training data. **This should be done using pipelines** and cross validation to choose your best model for each model type. You should compare your models using your metric chosen earlier.

Notes:

- You should set up a pipeline in `pyspark` for each of your models
- You should do your transformations using the functions from MLlib to easily put them into the pipeline. **At least one of the pipelines should use four or more transformations** prior to the model fit (`estimator`)
  - VectorAssembler counts as a transformation
  - Doing something like a log transform counts as well
  - Adding polynomial terms or interaction terms counts
  - etc.
- You can use the same set of transformations for multiple models (if appropriate)

Model Testing

Lastly, you should evaluate the best models from each class on the test set and state which overall model was deemed the best.