CS 5304 Assignment #1
Due February 20th, 2016

The goal of this assignment is to generate the equivalent of a Data Science "Hello World" application for Spark 2.0.2. This is your first practice for proving to the Instructor Team that you are capable of building a Data Analysis pipeline for Spark.

You will:

1. work with the Titantic Data Set from Kaggle at https://www.kaggle.com/c/titanic
2. build source code in Scala or Python that runs in Spark 2.0.2 to analyze the Titantic data set.
3. answer the question: "for subgroups of people boarding the Titanic, how would you maximize their individual probability of survival?". You must define meaningful subgroups. You should submit your predictions in a file that clearly labels identity of person and the prediction.
4. build at least two of {Naïve Bayes, Logistic Regression, random forests, support vector machines or neural networks using the libraries of Spark.MLLib only. Explain your choice; plot learning curves; explain observed behavior; investigate which features are most informative; do at least one round of error analysis to maximize your chosen metric (F1, accuracy, weighted F1); explain your choice of metric.
5. complete an analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy.
6. convey your analysis in writing and with supporting visualizations

Tier 1 Requirements:

1. Will answer or address every requirement in the instructions.
2. Submit working source code written in Python or Scala for Spark 2.0.2. The code will be modular, easy to read, contain informative comments, and will correctly conduct the analysis described in a maximum 2-page report.
3. As a maximum 2-page addendum to the report, the work will contain a statistical summary of the original data, a discussion of model convergence, a learning curve visualization, and a summary of the compute requirements for processing the data.

Tier 2 Requirements:

1. A maximum 2-page addendum to the report will provide visual evidence that the student used Spark running on Amazon Web Services (AWS) to complete the project, a summary stating the options for running Spark on AWS versus other cloud environments, and a discussion of the desirability for use of machine learning model types available in Spark.MLlib to complete the project.
2. A maximum 1-page addendum to the report that analyzes the value of each "column" or feature of the data available in the Titanic data set.

Tier 3 Requirements:

1. A maximum 1-page discussion of the value of this assignment. Why would we assign it? What value should you obtain from it? How does this exercise compare with selecting data science tools for a first project at a recently formed Startup company?
2. A maximum 1-page discussion with suggestions for how we could improve the assignment.
3. A maximum 2-page discussion describing how you will build a data science pipeline for analyzing large and small data sets for this course.