Final Group Project (10 % of grade)

Training from Disaster – This assignment will use the data from the Kaggle Titanic competition: https://www.kaggle.com/c/titanic/overview

The goal is to be able to predict whether a passenger is likely to survive the Titanic or not.

To do this exercise, we will have to do data cleaning to prepare the data. This means you may have to create a variety of variables to help learn to predict whether a given passenger on the Titanic was able to survive. There is a ton out on the web (including here) about this dataset, as it's popular among those just coming up to speed on machine learning classification models. After satisfying the project requirements, feel free to play around and use what you learned in class to join the Kaggle competition!

Project requirements:

1. Download the titanic data from Blackboard.

2. Below is the data dictionary for the data.

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg (France), Q = Queenstown (Ireland), S = Southampton (England)

3. Look at the data and see if there are any summary statistics that might give you some insights.

4. Data Cleaning and Feature Engineering

Sadly sci-kit learn will only let use numeric or boolean variables for analysis, so let's transform some of our variables to address that.

- Create booleans for each of the embarkment points.
- Create a boolean for is male.
- Create a boolean for whether someone has a cabin.

You may want to use One Hot Encoding for this.

Moreover, some of our ages are missing, so let's enter the missing values as 100 for now.

Some other feature engineering that you may want to perform can be found in this tutorial:

https://triangleinequality.wordpress.com/2013/09/08/basic-feature-engineering-with-the-titanic-data/

5. Creating an accurate model:

We have studied many ways to analyze data this semester. The titanic data is a classification problem.

- Divide the data into a training set and a test set.
- Pick three classification algorithms. Build a model for each of these methods.
- Train your algorithm on the training set. Test the model's accuracy on the test set
- Look at the accuracy metrics (as discussed in class) for each algorithm

Can you improve the model?

- Use regularization and/or ensemble methods to see if you can improve the accuracy of your model
- Use cross validation to see if that can improve the accuracy of the model.

Visualization – For the methods that you are able to show visualizations (i.e. decision trees), include these in your report.

Hand in: All python code.
An in depth analysis of your methodology that includes:
Why you chose the model.
Which models were more accurate?
How did you determine accuracy?
Methods to improve accuracy.
How did you test your model?

And anything else….

You will be graded on the completeness of your analysis. Think in terms of your company boss handing you a dataset and asking for an analysis of company performance. What would you need to say to keep your job???

*****Extra Credit (add 3 percentage points to your grade!! I asked you to apply three different machine learning algorithms to create a model for this dataset. Explain your process, with code, in a Jupyter notebook.

/