Daniel Hanks Jr
IST565 – Assignment Three

This labs purpose is to understand algorithms available to accomplish a classification task. The data is taken from the Titanic dataset found at https://www.kaggle.com/c/titanic/data, and is taken from the historical sinking of the Titanic. The data being analyzed deals with different classifications of people, such as gender, age, passenger class, etc. We'll then apply the model to predict who survived or not.

The data being used has been split into two groups, a training set and test set. The training set provides the outcome for each passenger and is used to generate predictions for the test set. Both datasets include PassengerId, Pclass, sex, age, SibSp, Parch, Ticket, Fare, Cabin and Embarked variables. The difference is the Training set includes the data on whether or not the passenger survived. Variable descriptions are as follows:

survival Survival
(0 = No; 1 = Yes)
pclass Passenger Class
(1 = 1st; 2 = 2nd; 3 = 3rd)
name Name
sex Sex
age Age (years)
sibsp Number of Siblings/Spouses Aboard
parch Number of Parents/Children Aboard
ticket Ticket Number
fare Passenger Fare
cabin Cabin
embarked Port of Embarkation

I will build naïve Bayes, kNN, SVM and Random Forest models to predict who survived or not, which will allow us to evaluate the different models. I'll be submitting the results of these models on to Kaggle who uses a variety of different error metrics to score the model. This will allow us to easily compare the models performance.

The first model I'm using is naïve Bayes. This classifier is actually a family of simple probabilistic classifiers based on applying Bayes' theorem. (sckikit-learn.org) I used the "RWeka" and the "e1071" library in R. These allowed me to run the naiveBayes algorithm on the two datasets that were created. A new spreadsheet was written that uses this algorithm to predict the survivors. An example of this spreadsheet are shown here:

	A	B
1	Passengerid	Survived
2	892	0
3	893	0
4	894	0
5	895	0
6	896	0
7	897	0
8	898	0
9	899	0
10	900	1

This turned out a Kaggle score of .74163.

The next model is the kNN or k-nearest neighbors algorithm. This algorithm stores all available cases and classifies new cases based on a similarity measure. Using the same datasets created in the previous example, I used the "class" library to run the kNN algorithm. The Kaggle score for this algorithm came in at .62201, which was pretty low compared to the naïve Bayes algorithm.

The next model is the SVM model. SVM stands for Support Vector Models. This model also requires the e1071 library in R. This model basically takes classified data (training data) and makes a prediction with it. The score for this algorithm was .77990, giving us the best score thus far.

The final model is the random forest model. A random forest is an ensemble of decision trees which output a prediction value. The decision trees are constructed by using a random subset of the training data. You train your forest and then pass each test row trhough it in order to output a prediction. The randomForest package used in R allows us to print the dataset showing the error rate and the confusion matrix, as shown here:

```
Call:
 randomForest(formula = Survived ~ ., data = newtrain, ntree = 10)
               Type of random forest: classification
                     Number of trees: 10
No. of variables tried at each split: 2

        OOB estimate of  error rate: 19.24%
Confusion matrix:
    0   1 class.error
0 478  70   0.1277372
1 101 240   0.2961877
```

The random forest model had the same score as the SVM model with .77990.

SUMMARY

MODEL	KAGGLE SCORE		
naïve Bayes	.74163		
kNN	.62201		
SVM	.77990		
Random Forest	.77990		

Pic of Kaggle leaderboard:

2834	new	OksanaPazdriy	0.77990	1	Mon, 17 Oct 2016 14:52:48
2835	new	sajal 2	0.77990	3	Mon, 17 Oct 2016 19:42:48
2836	new	TimonSchmelzer	0.77990	3	Mon, 17 Oct 2016 21:36:09 (-0.2h)
2837	new	**Dan Hanks Jr**	**0.77990**	**4**	**Tue, 18 Oct 2016 01:00:36 (-1h)**
2838	new	Vic Vijayakumar	0.77990	3	Tue, 18 Oct 2016 02:40:23
2839	new	Lokendra Singh	0.77990	1	Tue, 18 Oct 2016 08:27:31
2840	new	Vitaly Sh	0.77990	1	Tue, 18 Oct 2016 10:37:40

References:

http://scikit-learn.org/stable/modules/naive_bayes.html

https://www.kaggle.com/c/titanic/details/new-getting-started-with-r