Disaster Relief Project

In this project, you will use classification methods covered in this course to solve a real historical data-mining problem, locating displaced persons living in makeshift shelters following the destruction of the earthquake in Haiti in 2010.

Following that earthquake, rescue workers, mostly from the United States military, needed to get food and water to the displaced persons. But with destroyed communications, impassable roads, and thousands of square miles, locating the people was critical.

As part of the rescue effort, a team from the Rochester Institute of Technology were flying an aircraft to collect high resolution geo-referenced imagery. It was known that the people whose homes had been destroyed by the earthquake were creating temporary shelters using blue tarps, and these blue tarps would be good indicators for locating the people—if only they could be located in time, out of the thousands of images that would be collected every day. The problem was that there was no way for people to search the thousands of images in time to find the tarps and communicate the locations back to the rescue workers in time. The solution would be provided by data-mining algorithms, which could search the images far faster and more thoroughly than humanly possible. The goal was to find the best algorithm that could search the images and locate displaced persons in time for the locations to be communicated back to the rescue workers.

This disaster relief project is the subject matter for your project in this course, which you will submit in two parts. You will use data from the actual collect over Haiti. Your goal is to test each of the algorithms you learn in this course on data from the Haiti imagery and determine which works most accurately and in a timely way to try to locate as many of the displaced persons in imagery so they can be provided food and water in time.

You will document the accuracy and runtime for classification for each algorithm in the Disaster Relief Project Powerpoint files. In Module 6, you will submit your Powerpoint file with some of the algorithm results filled out. In Module 12, at the end of the course, you will submit your completed file, which will include your conclusions from your research and recommendations for what algorithms should be used.

Project Part 1 (60 points)

The Powerpoint file you will submit will have the following sections.

Performance Metrics, 10x cross-validation:

Method	**Accuracy**
KNN (K=)	
LDA	
QDA	
Logistic Regression	

Results (30 points)
Fill in the % accuracy you get for each method **(24 points; 0, 3, or 6 points each based on reasonableness of answer)**
Fill in the value you used for K in KNN **(0, 3, or 6 points based on reasonableness of answer)**
For a **2-point** bonus, add another column next to Accuracy and include the AUC (area under the curve) values for each method.
For an additional **2-point** bonus, add another slide with ROC curves for all methods.

Conclusions (30 points)
State at least three conclusions.
One of your conclusions must be your determination of which algorithm works best. (**10 points total; 0, 5, or 10 points each based on reasonableness of answer)**
Your other conclusions should be observations you make based on the work on this project, such as the following: (**0–20 points based on the quality of the conclusions)**

- Were there multiple adequate performing methods or just one clear best method?
- Are there actions you recommend that might improve results?
- Do these data seem particularly well-suited to one class of prediction methods, and if so, why?

Project Part 2 (142 points)
The Powerpoint file you will submit will have the following sections.

Performance Metrics, 10× cross-validation:

Method	**Accuracy**
KNN (K =)	
LDA	
QDA	
Logistic Regression	
Random Forest (parameters =)	
SVM (kernel type =)	

Results: (36 points)
Fill in the % accuracy you get for each method. **(24 points total; 0, 2, or 4 points each based on reasonableness of answer)**
Fill in the parameter value(s) you used in each of the following methods: **(12 points total; 0, 2, or 4 points each based on reasonableness of answer)**

- Value you used for K in KNN
- Your parameters for random forest
- The kernel you used for SVM

For a **3-point** bonus, add another column next to Accuracy and include AUC (area under the curve) values for each method.

Performance Metrics, Hold-Out Data:

Method	**Accuracy**
KNN (K=)	
LDA	
QDA	
Logistic Regression	
Random Forest (parameters =)	
SVM (kernel type =)	

Results: (36 points)
Fill in the % accuracy you get for each method. **(24 points total; 0, 2, or 4 points each based on reasonableness of answer)**
Fill in the parameter value(s) you used in each of the following methods: **(12 points total, 0, 2, or 4 points each based on reasonableness of answer)**

- Value you used for K in KNN
- Your parameters for random forest
- The kernel you used for SVM

For a **3-point** bonus, add another column next to Accuracy and include AUC (area under the curve) values for each method.

Conclusions: (60 points)
State at least six conclusions.
In your conclusions, you must explain the following: (**0–10 points based on the quality of the conclusions)**

- Which algorithm works best in the cross-validation data
- Which works best on the Hold-Out data
- Which algorithm you recommend for use in detection of blue tarps