Project 2-- Cloud Data

Tiantian Fu:3033299999
Grace Wang:3033252224

0. Acknowledgement

Contributions: Grace did the paper summary, most of the coding part and the explanations in Q4 diagnostics. Tiantian generated the graphs, wrote the observations and the report and did Q5 reproducibility part. We switched and checked each others work.

Proceed the project: we basically checked the lecture slides and notes to review different regression models and relevant materials. As we got stuck on the coding part, we checked google and we especially thanks Shifan Jin to help us debug some issues. We put all the libraries we used in R in the coding part.

1. Data Collection and Exploration

a) Summary:

Global climate change caused by carbon dioxide is always a heated topic. Climate models could predict the strong dependencies of surface air temp on increasing atmospheric carbon dioxide levels in Arctic and scientists need to check those dependencies. To do more investigation, we need to use cloud detection and measure the cloud coverage across the Arctic region.

The article uses multiangle imaging spectroRadiometer(MISR) imagery onboard the NASA Terra satellite which has nine cameras to view the earth at different angles, so that it could process the massive MISR data and scientists could do further exploratory analysis to construct a labeling scheme. The data used were collected from 10 MISR orbits of path 26 over the Arctic, northern Greenland, and Baffin Bay. And there are 57 data units contained with 7,114,248 1.1-km resolution pixels with 36 radiation measurements for each pixel in the study. Around 71.5% of the valid pixels labeled “clear” or “cloudy” by experts in order to evaluate the algorithms. The first algorithm is called enhanced linear correlation matching(ELCM), which thresholds the features with either fixed or data-adaptive cutoff values. The second algorithm is ELCM-QDA, which uses the labels resulted from the ELCM algorithm to produce the probability labels of “cloudiness” as a more informative probability prediction. To evaluate the accuracies of the two algorithms above, scientists use them to apply to extensive MISR data sets and tested against expert labels.

Scientists demonstrated that three physical features and they contain contain sufficient information to separate clouds from ice- and snow-covered surfaces. The ELCM algorithm based on the three features is more accurate and provides better spatial coverage than the existing MISR operational algorithms.

This cloud research improved understanding of the flow of visible and infrared radiation through the atmosphere, so scientists can tease apart the response of clouds to changes in arctic

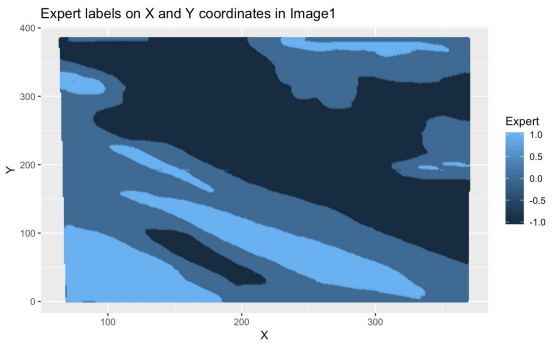
climate and their feedback on it. This work has positive impacts for statistics that go beyond technical development and implementation of statistical methods. It also demonstrates the power of statistical thinking, and also the ability of statistics to contribute solutions to modern scientific problems.

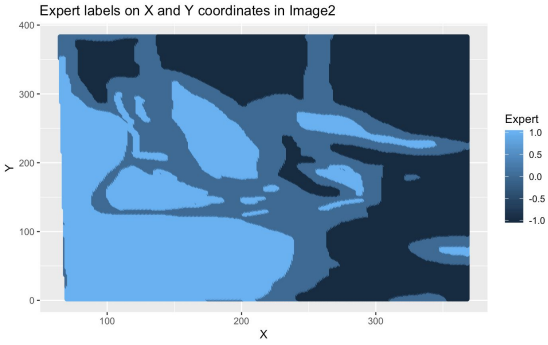
b) Data Summary

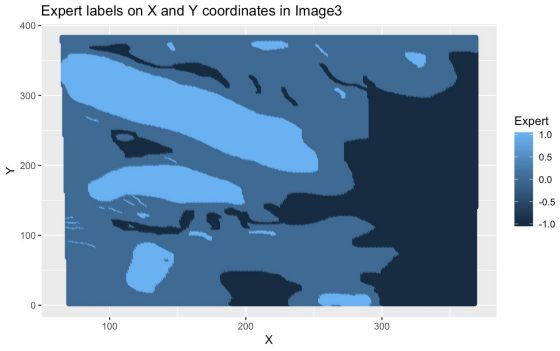
By summarizing the data, we got the percentage of pixels for different classes.

	cloud	unlabeled	clear
image1	0.1776549	0.3845560	0.4377891
image2	0.3411172	0.2863522	0.3725306
image3	0.1843825	0.5226746	0.2929429

In these three images, around 70% of the pixel labels are valid(clear or cloud)





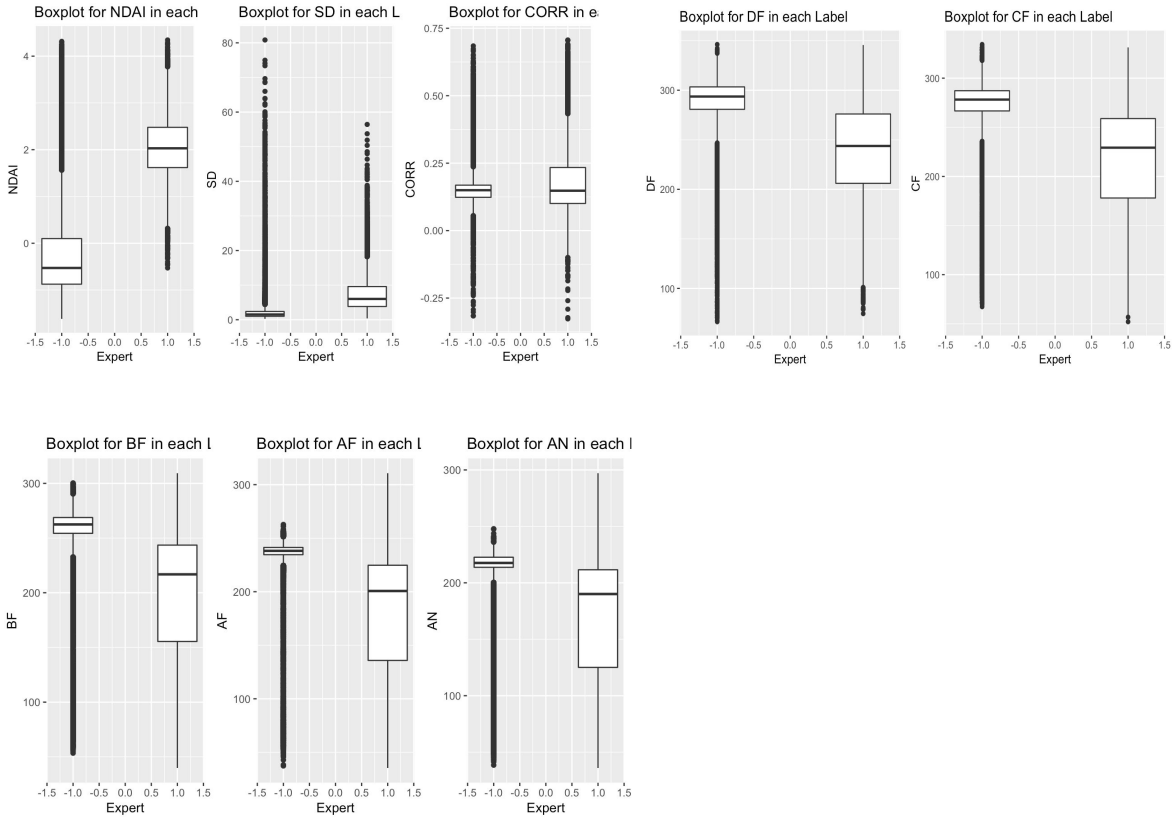


Pattern/Trend: By looking at these three graphs we can tell that the data points that have the same label(cloud,no cloud and unlabeled) tended be clustered together.

The reason the assumption is not i.i.d is because the points that have the same label are gathered together, which means the whole datasets are not randomized.

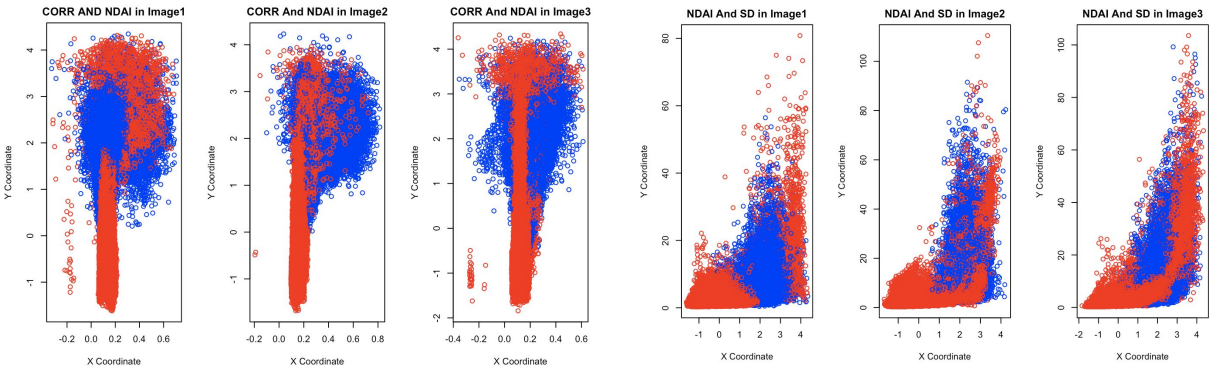
c) EDA

We used boxplots to see the relationships between expert labels and individual features.



Two classes(cloud,no cloud) behave differently based on the radiance. As we could see in the five boxplots, no cloud data have higher radiance in different angles(DF,CF,BF,AF,AN)

We used scatter plots to summarize pairwise relationship between the features themselves.



From the scatter plots for each two features in every image, we can see that the graphs for any two features do not have strong

positive or negative relationships with each other. Therefore, there is no collinearity between each two features.

2. Preparation

a)Data Split

We used two different ways to split the data.

First of all, we divided each image into 5x5 blocks by x,y coordinates of each observation and then, we randomly sampled training blocks, validation blocks and testing blocks proportional to ⅗, ⅕, ⅕ of the 25 blocks respectively in each image. Secondly, we combined training set, validation set and test set from each image and got the whole training set, validation set and test set.

The second way we used was that we sampled ⅗ of the observations in each image as training set, ⅕ of the observations in each image as test dataset, and ⅕ of the observations in image as validation dataset. And then we combined all training sets, validation sets and testing sets from three images.

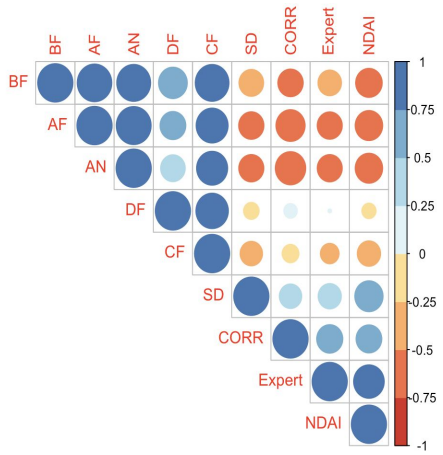
b)Baseline

We set the trivial classifier by setting all labels to -1(cloud-free) on the validation set and on the test set generated from the first splitting method. By comparing the true labels on the validation set and on the test set with the trivial classifier, we got the accuracies 75.14% and 60.14% respectively on validation set and test set.

And we got the accuracies 61.119% and 61.003% respectively on validation set and test set that generated from the second splitting method.

If our validation set and test set tended to have observations of cloud-free areas, the trivial classifier will have a high average accuracy.

C)First Order Importance



	Expert	NDAI	SD	CORR	DF	CF	BF	AF	AN
Expert	1.0000000	0.7584104	0.4360264	0.5510043	0.0107873	-0.2827573	-0.4476648	-0.5073210	-0.5045998
NDAI	0.7584104	1.0000000	0.6474474	0.5350207	-0.1639961	-0.4384729	-0.5710109	-0.6119935	-0.6085254
SD	0.4360264	0.6474474	1.0000000	0.4073057	-0.1965740	-0.4070290	-0.4912370	-0.5143340	-0.5068789
CORR	0.5510043	0.5350207	0.4073057	1.0000000	0.1477618	-0.2290945	-0.5182108	-0.6840183	-0.7460751
DF	0.0107873	-0.1639961	-0.1965740	0.1477618	1.0000000	0.8503037	0.6703445	0.5377937	0.4892642
CF	-0.2827573	-0.4384729	-0.4070290	-0.2290945	0.8503037	1.0000000	0.9189584	0.8259473	0.7795202
BF	-0.4476648	-0.5710109	-0.4912370	-0.5182108	0.6703445	0.9189584	1.0000000	0.9624793	0.9255600
AF	-0.5073210	-0.6119935	-0.5143340	-0.6840183	0.5377937	0.8259473	0.9624793	1.0000000	0.9819174
AN	-0.5045998	-0.6085254	-0.5068789	-0.7460751	0.4892642	0.7795202	0.9255600	0.9819174	1.0000000

By using correlation function and plotting the correlation between two variables, we can tell that NDAI,CORR,AF has the highest coefficients among all variables. However, if we take a closer look at the graph, we can notice that SD has collinearity between both AN and AF. Therefore, we have decided to use NDAI, CORR, and SD as our best three features.

d)Generic Cross Validation Function
See github folder.

3. Modeling

a) Report accuracies for different classification methods
Results from Dataset1 (generated from the first way of splitting data)

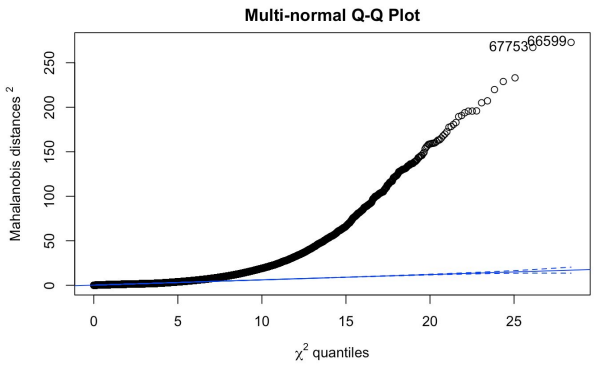
	LDA	QDA	KNN	GLM
fold1	0.8507246	0.8589521	0.8795488	0.8521424
fold2	0.8516383	0.8620309	0.8836447	0.8490233
fold3	0.8542894	0.8573679	0.8798954	0.8499008
fold4	0.8521693	0.8632640	0.8802495	0.8511704
fold5	0.8550994	0.8607435	0.8821398	0.8496172
Average Mean	0.8527842	0.8604717	0.8810957	0.8503708

Results from Dataset2 (generated from the second way of splitting data)

	LDA	QDA	KNN	GLM
fold1	0.8992490	0.8942926	0.9126464	0.8924001
fold2	0.8982878	0.8976570	0.9074797	0.8904175
fold3	0.8966056	0.8990988	0.9088315	0.8922467
fold4	0.8956143	0.8971763	0.9089817	0.8948333
fold5	0.8987384	0.8975068	0.9118053	0.8933646
Average Mean	0.8976990	0.8971463	0.9099489	0.8926524

We have used four different classification methods including LDA,QDA,KNN and GLM for both datasets generated from two different splitting methods. By using the function we created from 2(d), we got the accuracies(1-CVgeneric) and CVgeneric output 0-1 loss function's rate. We got the test accuracies and the average across each fold.

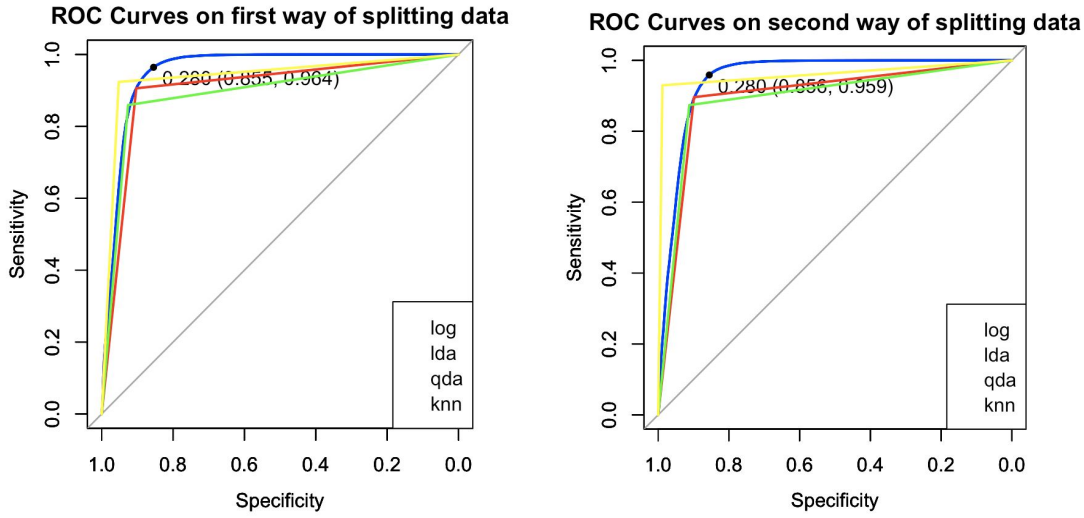
Result commentary: We found that the KNN performs best among the methods since it has the highest average accuracies for both splitting ways.



Assumption 1: we assumed that in lda, qda methods, predictor variables X are drawn from multivariate Gaussian distribution. By plotting the Multi-Normal QQ plot, we found that the mahalanobis distance is a little bit away from the blue line as the quantile increases, which means that the X data does not really follow multi-normal distribution. This is a problem during the research project.

Assumption 2: In those modelings we used, we assumed that only NDAI, CORR, and SD will be the predictor variables since we chose them as the “best” three features based on the proof in the last problem.

b) ROC Curves for different classification methods on First Dataset



AUC values in split 1 data:

```
> log_roc$auc
Area under the curve: 0.9646
> lda_roc$auc
Area under the curve: 0.9208
> qda_roc$auc
Area under the curve: 0.9132
> knn_roc$auc
Area under the curve: 0.8673
```

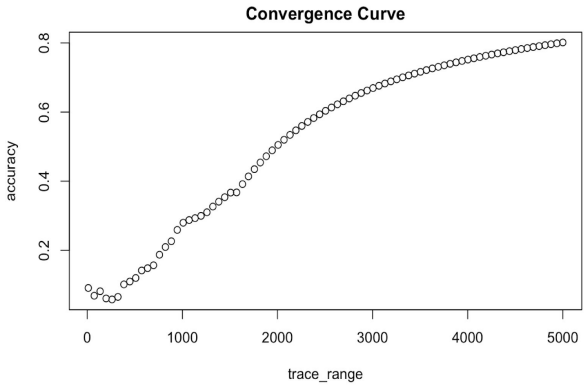
The reason that we chose cutoff value is that: sensitivity(True Positive Rate) and specificity(False Positive Rate) are both important to determine the cut off point and we want the point with a large sensitivity and low specificity(or say high 1-specificity).
From the graph we could see that the ROC curve for logistic regression has highest AUC (Area under the curve), which is around 0.96 thus we pick the cutoff point from the blue curve(logistic regression) and choose a "upper left corner" which approximately around(0.858,0.953).

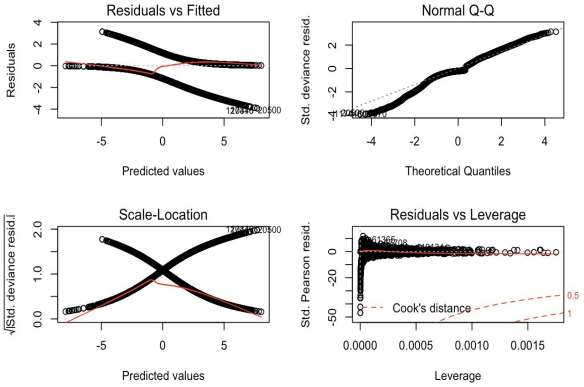
4. Diagnostics

a)An In-depth Analysis of a Good Classification Model

Firstly, we used the first method of splitting data and chose the glm model. And we tried different number of test sets and produced predictions in order to check the accuracies with various sample sizes. The graph shows below, and we can see that as the size becomes larger, the accuracy goes to converge to a relatively steady state which is around 0.80.

We also diagnosing the regression model by checking the four plots on the right side below. As we can see, in Residuals vs Fitted, the residuals are not spreaded randomly around the zero line, which shows heteroskedasticity; in Normal QQ, the points on QQ are mostly following the straight diagonal line, but some points are below the straight line on the left side, and some points are above the line on the right side, thus the tails of the normal distribution are long.





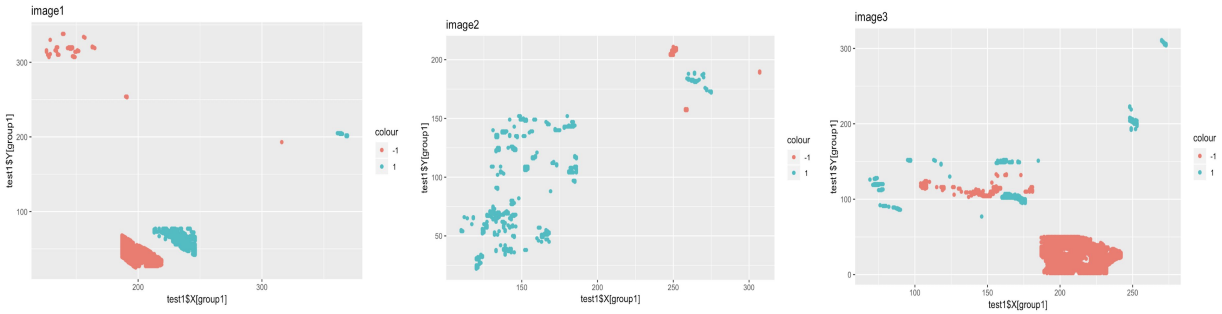
b)Patterns in Misclassification Errors

Our best classification model is logistics regression. We used the training set, validation set and test set generated from the first method of splitting data to get misclassification graphs in each image.

By looking at the X,Y coordinates distribution of the misclassification data, we can see that in image1, there are not many misclassified points in the graph except in the area where X's are in the range from 175 to 250 and Y's are in the range from 0 to 100. And the two groups (label1 and label-1) are separately formed. In terms of Image2, We see that the most misclassified points are labeled as one (blue points) and they are spreaded out. In terms of Image3, there are both blue points and red points occurred in the graph, and there is a cluster of red points formed in the area where X's are in the range from 180 to 240 and Y's are in the range from 0 to 50.

By looking at the histograms of all test data vs. misclassification data for each feature in each image, we could see that three different images have different specific range of misclassification data. In image1, the range of NDAI values are specified from 1.1 to 4.3, CORR values are specified from 0 to 0.55, SD are specified from 0 to 60. In image2, the NDAI values are mostly located around 1.4 to 1.6 and 1.8 to 2.1 , CORR values are spread out from 0.12 to 0.17, and SD are ranged from 0 to 25. In image3, the NDAI values are mostly located around 1.5 to 4.5, CORR values are mostly located around 0 to 0.3, and SD are mostly located around 0 to 50.

X,Y coordinates distribution of the misclassification data in different Images:



Feature Values Distributions Comparisons between all data and misclassification data:
image1:

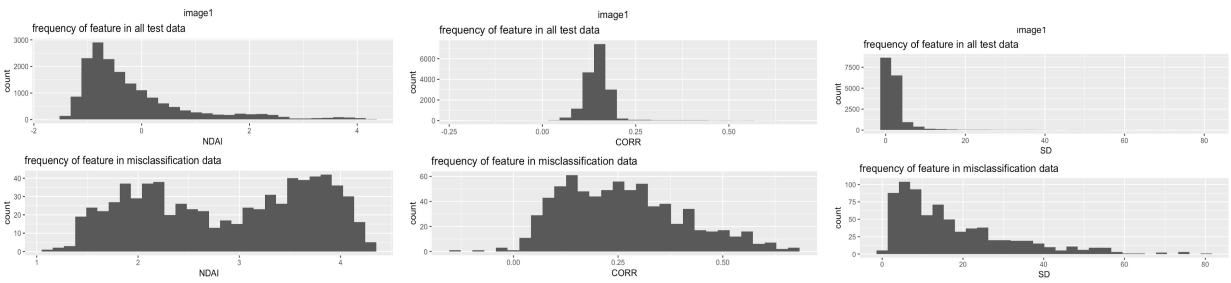


image2:

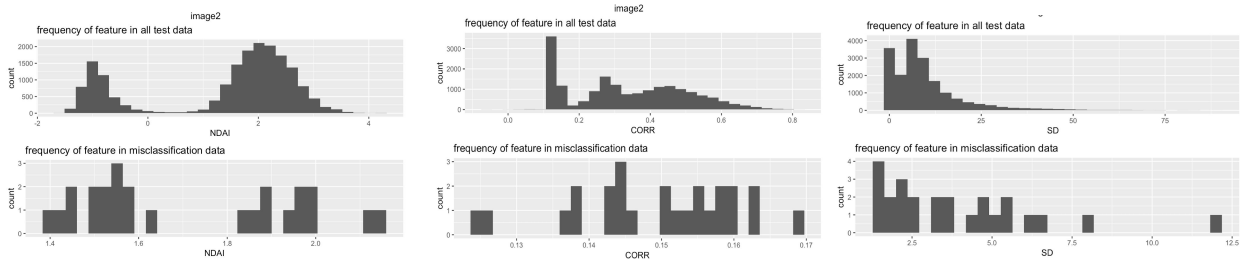
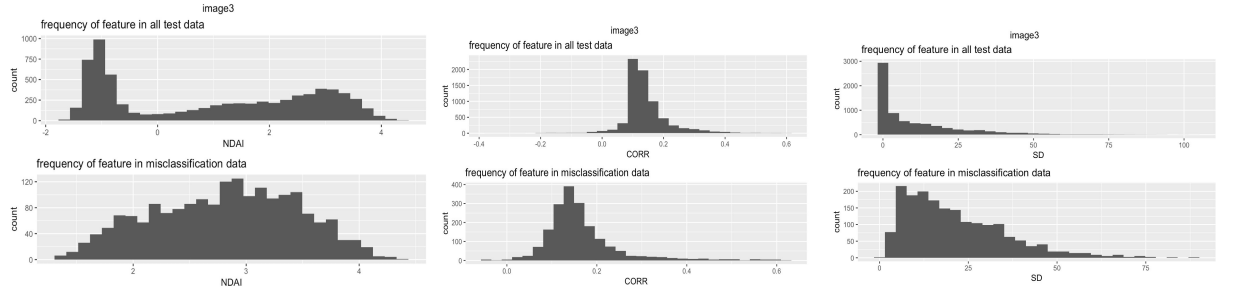
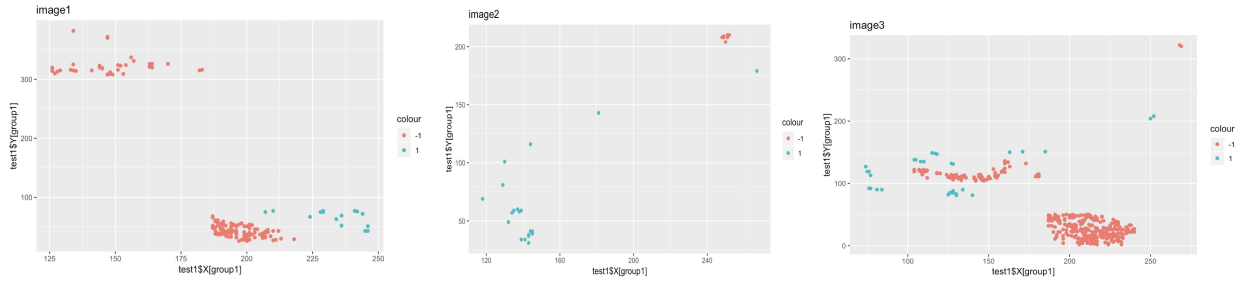


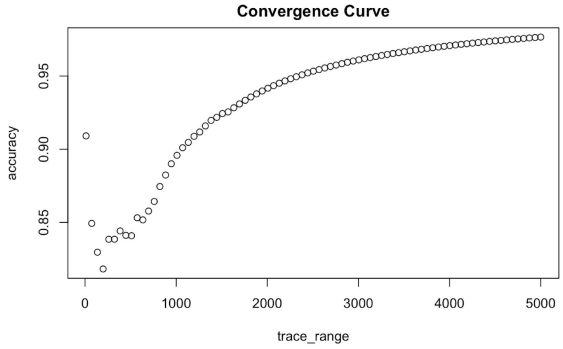
image3:



c)



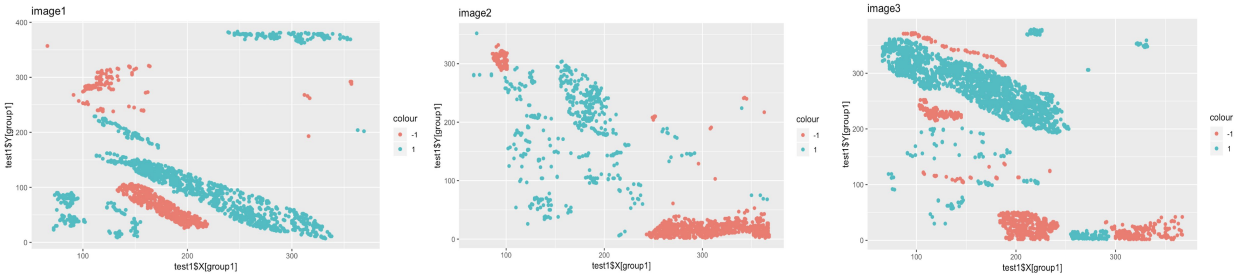
By using random forest as classifier, since the three misclassification plots has a big improvement comparing to the previous plots from 4(b). In the three plots, the amount of data that is shown misclassified has decreased a lot.In image1, the misclassified data are mostly the -1 colour group located around (X = 185 ~ 215, Y = 50) and (X = 125 ~ 180, Y = 325). In image2, there is barely data that is misclassified for both colour. In image3, although the amount of misclassification has decreased, there is still some data remained misclassified for both colour. For colour -1, there is a certain amount of data around (X = 180 ~ 240, Y = 30) and (X = 100 ~ 175, Y = 110) that is misclassified. For colour 1, there is only a few data located around (X = 50 ~ 175, Y = 100) that is misclassified.



And also by looking at the convergence curve, we can see that the accuracy rate also improved from 80% to 98%.

In order to see how well the model will work on future data without expert labels, we added the random forest classification model into the generic function and we calculated the loss rate for using random forest classification and it turned out we had almost 100% accuracy rate. Therefore, we think this model will perform pretty well on future data without expert labels.

d)



In the three misclassification plots, we can see that the amount of misclassified data has increased a lot for both colour 1 and colour -1. In addition, there seems to be no specific patterns in the location of X and Y since we can see from the plots that they are all located randomly, which corresponds to the second way of splitting data (randomly).

image1:

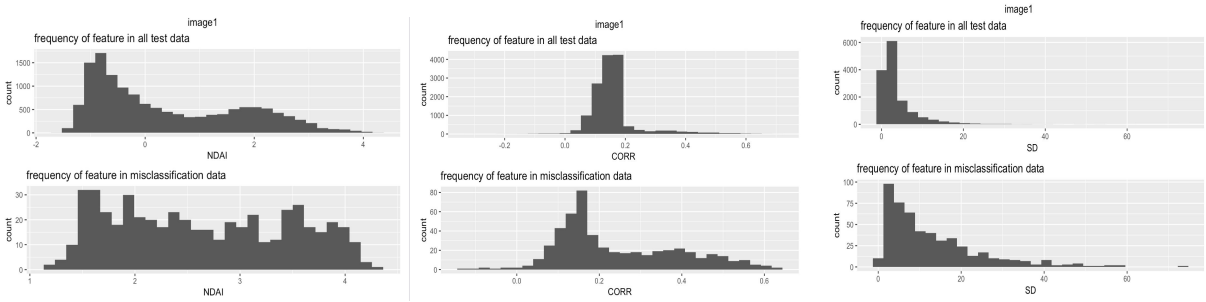


image2:

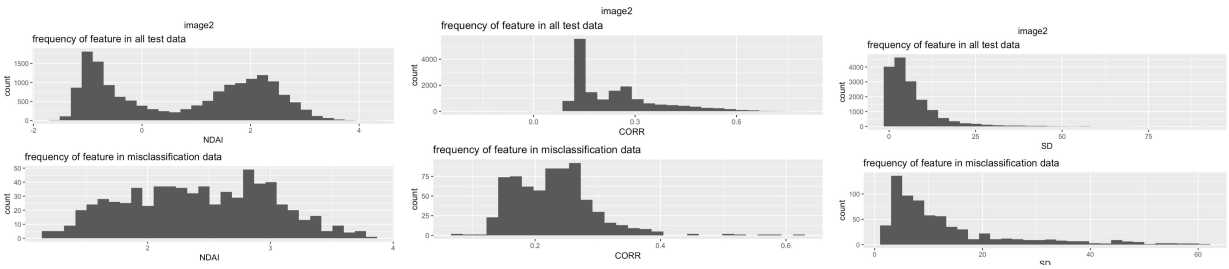


image3:

The results in part 4(a) and 4(b) do change as we change the splitting data method to the random way (the second way of splitting data), the first observation that comes to us is that for each image, the distribution of the misclassification data in the X,Y coordinates becomes messy and more spreaded. As for three features(NDAI,AF and CORR), the ranges of wrong labels also become more spreaded and don't have any specific ranges among the whole test data.

e)Conclusion:
In the process of identifying the best classification method, we have the following candidate methods: logistic regression, linear/ quadratic discriminant analysis , K-nearest neighbors.

We used different methods such as ROC curves, misclassification plots, convergence curves to find the best classification model. And it turned out that logistic regression is the best model since it has largest AUC area for both splitting methods.

However, on the basis misclassification plots and histograms of corresponded features in 4(a) and 4(b), we think that random forest classification is a better classifier compared to logistics method because random decision forests correct for overfitting. And by looking at the misclassification plots and convergence curve generated from random forest method, we verified it as our better classifier. Lastly, we fit the other data set from method two of splitting data into logistics regression classifier and we found out the results are worse than the first dataset(generated from the first splitting method).

5. Reproducibility

Github link: https://github.com/TiantianFu/STAT-154-Project-2-Cloud-Data