Frank Chen 26823851
Xiaoyi Xu 3034505111
1.(a) A) Purpose of the Study: The sensitivity of the earth's climate to the increasing amount of Carbon Dioxide is among the general interest of scientific field. The climate models suggest that the strongest dependencies of surface air temperatures on increasing atmospheric CO2 will happen in Arctic. Determining the properties of cloud is challenging due to the similarity of ice-water-particle cloud and ice-snow- surface particle. Despite the development of MISR method, the dataset is so large that traditional standard classification framework does not label the cloud suitably. Thus the goal of the paper is to build operational cloud detection algorithms that can efficiently process the massive MISR data set one data unit at a time without requiring human intervention.

B) Data: The data used in this study is from 10 MISR orbits of path 26 over the Arctic, northern Greenland, and Baffin Bay. The repeat time between 2 consecutive orbits over the same path was 16 days, so the 10 orbits span approximately 144 days from April 28 through September 19, 2002. The paper chooses path 26 for study because of the richness of its surface features. 6 data units from each orbit are included in this study. To evaluate the performance of the method and existing MISR algorithm, the paper uses manual labeling due to the rareness of high-quality ground-based measurement. After expert labeling, tools are used to label the pixels in MISR nadir camera images as clear or cloudy. The labels are used to evaluate the performance of different algorithms.

C) Collection Method:

1). Only MISR red radiation data are used for constructing features due to its having the highest spatial resolution of MISR. Overall data units, the paper investigated the distributions of a large collection of features including linear combinations of angular radiances, correlations among different angles and wavelengths, nonlinear transformations of radiances, spatial patterns of clouds, and smoothness of reflecting surfaces. Three physical features are found to differentiate surface pixels from cloudy ones. The first feature is an average linear correlation of radiation measurements at different view angles (CORR). High values of CORR suggest either clear (cloud-free) conditions or the presence of low altitude cloud that is registered to the same location on the underlying surface. To avoid errors in misclassification, second feature as SD (standard deviation within groups of MISR An camera red radiation measurements) and third feature (NDAI, average radiation measurements) . Then the paper labels pixels by thresholding the features, which is called ELCM algorithm. Based on results from different thresholds to the expert labels, CORR and SD are stable and robust across all data units, thus we set CORR = 0.75 and SD = 2.0. Clustering algorithm is used on current data unit and the threshold learned from previous data units at the same location to set threshold NDAI. Because the thresholds produced by the algorithm is not perfect, neither are the labels. Therefore, reporting a probability of cloudiness is desirable and more informative than providing only a binary clear vs cloudy model. Fisher's QDA trained from ELCM labels is used to provide an estimate of probability, or confidence, of cloudiness.

D) Conclusion and Potential Impact: The results, given in Table 1, show that the ELCM Algorithm agreement rate of 91.80% over the 5 million testing pixels is 8.57% higher than the MISR ASCM (83.23%) algorithm and 11.90% higher than the SDCM (80.00%) algorithm. This represents a significant improvement from both scientific and statistical standpoints. The offline SVM has an 80.99% agreement rate when using the expert labels for training, much lower than the ELCM algorithm (but comparable to the SDCM or ASCM algorithms). Although ELCM-QDA doesn't improve overall agreement rates, it goes beyond ELCM's binary labels of cloudy versus clear by providing probability labels.

Couple significant impacts include: 1). The explosion of earth science data supports statistians with a major role in data processing and getting live results. 2). It demonstrates the power of statistical thinking and the ability for statistics to solve modern science problems.

1.(b)As can be seen below, in image 1, there are 43.79% pixels for the unclouded class, 38.46% pixels for the unlabeled class and 17.77% pixels for the clouded class. In image 2, there are 37.25% pixels for the unclouded class, 28.64% pixels for the unlabeled class and 34.11% pixels for the clouded class. In image 3, there are 29.29% pixels for the unclouded class, 52.27% pixels for the unlabeled class and 18.44% pixels for the clouded class.

```
I1_prop = {-1: I1.groupby("expertlabel").count()["x"][-1.0]/sum(I1.groupby("expertlabel").count()["x"]), 0: I1.groupby("exper
I1_prop
#Thus from image 1 we have 43.8% of not cloud, 38.5% of unlabeled and 17.8% of cloud
```
```
Out[3]: {-1: 0.4377890982304802, 0: 0.38455597115309514, 1: 0.17765493061642468}
```
```
I2_prop = {-1: I2.groupby("expertlabel").count()["x"][-1.0]/sum(I2.groupby("expertlabel").count()["x"]), 0: I2.groupby("exper
I2_prop
#Thus from image 2 we have 37.3% of not cloud, 28.6% of unlabeled and 34.1% of cloud
```
```
Out[4]: {-1: 0.37253062288246025, 0: 0.28635218486666403, 1: 0.34111719225089043}
```
```
I3_prop = {-1: I3.groupby("expertlabel").count()["x"][-1.0]/sum(I3.groupby("expertlabel").count()["x"]), 0: I3.groupby("exper
I3_prop
#Thus from image 3 we have 29.3% of not cloud, 52.3% of unlabeled and 18.4% of cloud
```
```
Out[8]: {-1: 0.29294288169280575, 0: 0.5226746053099803, 1: 0.18438251299721395}
```

The images below are beautiful maps for the three images. Specifically, red region stands for clouded class, blue region stands for unlabeled class and green region stands for unclouded class.

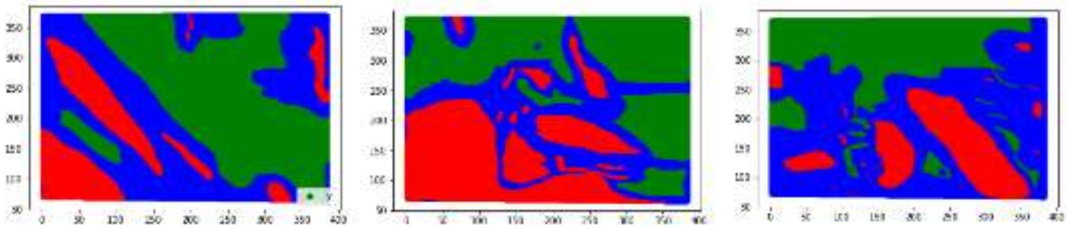


Image1 image2 image3

First, unclouded class mainly takes up the upper area of the image, while the clouded class mainly takes up the lower area of the image. Second, both unclouded class and clouded class appear like the cluster, namely, the near data are dependent, so we think an i.i.d. assumption for the samples can not be hold for this dataset .

1.(c) First, let's summarize pairwise relationship between the features themselves by analyzing feature correlation tables and pair-plots below.

In three images, NDAI, SD are highly correlated and DF, CF, BF, AF, AN are highly correlated. Then, let's summarize the relationship between the expert labels with the individual features. In three images, the correlation between NDAI and expertlabel is the highest compared to the correlation between other features and experlabel.

	expertlabel	NDAI	SD	CORR	DF	CF	BF	AF	AN
expertlabel	1.000000	0.659129	0.332461	0.144806	-0.427777	-0.439946	-0.438748	-0.415335	-0.383833
NDAI	0.659129	1.000000	0.601305	0.251090	-0.553169	-0.594024	-0.601799	-0.579128	-0.544755
SD	0.332461	0.601305	1.000000	0.165031	-0.525445	-0.526744	-0.517829	-0.498210	-0.470594
CORR	0.144806	0.251090	0.165031	1.000000	-0.235458	-0.433126	-0.566343	-0.668080	-0.732222
DF	-0.427777	-0.553169	-0.525445	-0.235458	1.000000	0.942325	0.893021	0.847884	0.808954
CF	-0.439946	-0.594024	-0.526744	-0.433126	0.942325	1.000000	0.968870	0.929755	0.893032
BF	-0.438748	-0.601799	-0.517829	-0.566343	0.893021	0.968870	1.000000	0.978946	0.946295
AF	-0.415335	-0.579128	-0.498210	-0.668080	0.847884	0.929755	0.978946	1.000000	0.983469
AN	-0.383833	-0.544755	-0.470594	-0.732222	0.808954	0.893032	0.946295	0.983469	1.000000

Image1

	expertlabel	NDAI	SD	CORR	DF	CF	BF	AF	AN
expertlabel	1.000000	0.682538	0.350987	0.692268	0.260880	-0.217409	-0.459476	-0.525865	-0.516722
NDAI	0.682538	1.000000	0.629533	0.556630	0.077767	-0.319146	-0.478830	-0.501901	-0.494904
SD	0.350987	0.629533	1.000000	0.342583	-0.172349	-0.420468	-0.459183	-0.443164	-0.431128
CORR	0.692268	0.556630	0.342583	1.000000	-0.015108	-0.492139	-0.740004	-0.835610	-0.872558
DF	0.260880	0.077767	-0.172349	-0.015108	1.000000	0.720627	0.498306	0.409894	0.396417
CF	-0.217409	-0.319146	-0.420468	-0.492139	0.720627	1.000000	0.884021	0.814971	0.793740
BF	-0.459476	-0.478830	-0.459183	-0.740004	0.498306	0.884021	1.000000	0.957910	0.929749
AF	-0.525865	-0.501901	-0.443164	-0.835610	0.409894	0.814971	0.957910	1.000000	0.974852
AN	-0.516722	-0.494904	-0.431128	-0.872558	0.396417	0.793740	0.929749	0.974852	1.000000

Image2

	expertlabel	NDAI	SD	CORR	DF	CF	BF	AF	AN
expertlabel	1.000000	0.498879	0.235972	0.342745	0.142414	0.021682	-0.057226	-0.128419	-0.172904
NDAI	0.498879	1.000000	0.675159	0.408663	0.008829	-0.155502	-0.270672	-0.360589	-0.398935
SD	0.235972	0.675159	1.000000	0.367638	-0.052295	-0.244918	-0.377850	-0.445954	-0.456228
CORR	0.342745	0.408663	0.367638	1.000000	0.306289	0.162637	-0.119208	-0.376970	-0.520368
DF	0.142414	0.008829	-0.052295	0.306289	1.000000	0.796753	0.646610	0.531275	0.466156
CF	0.021682	-0.155502	-0.244918	0.162637	0.796753	1.000000	0.863213	0.723591	0.633737
BF	-0.057226	-0.270672	-0.377850	-0.119208	0.646610	0.863213	1.000000	0.907953	0.804141
AF	-0.128419	-0.360589	-0.445954	-0.376970	0.531275	0.723591	0.907953	1.000000	0.936600
AN	-0.172904	-0.398935	-0.456228	-0.520368	0.466156	0.633737	0.804141	0.936600	1.000000

Image3

Image1

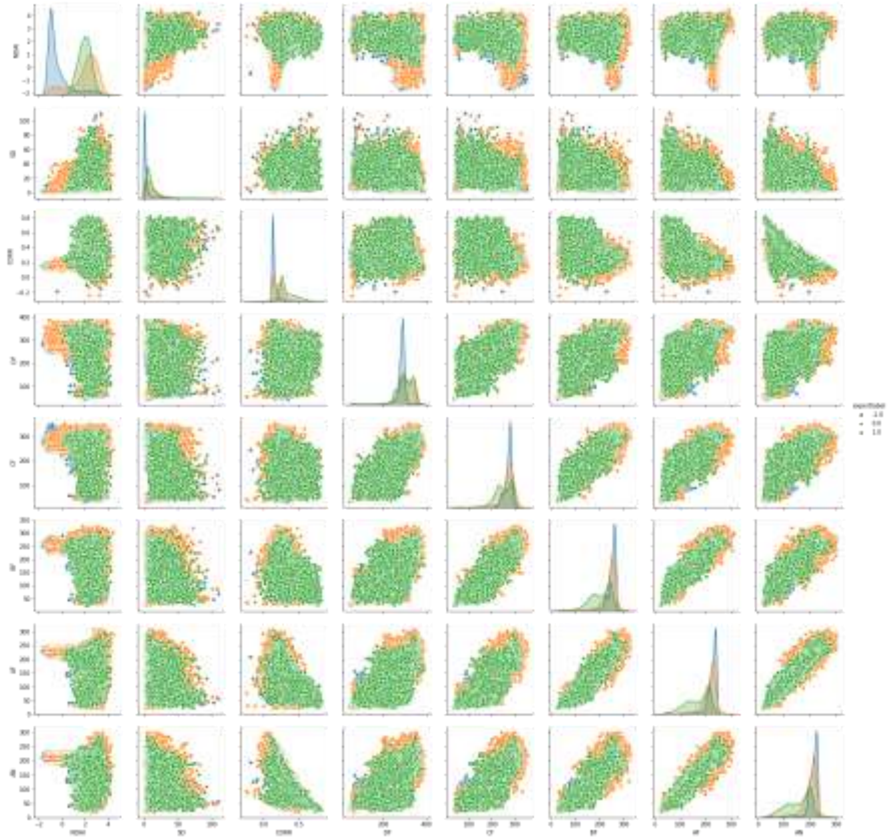


Image2

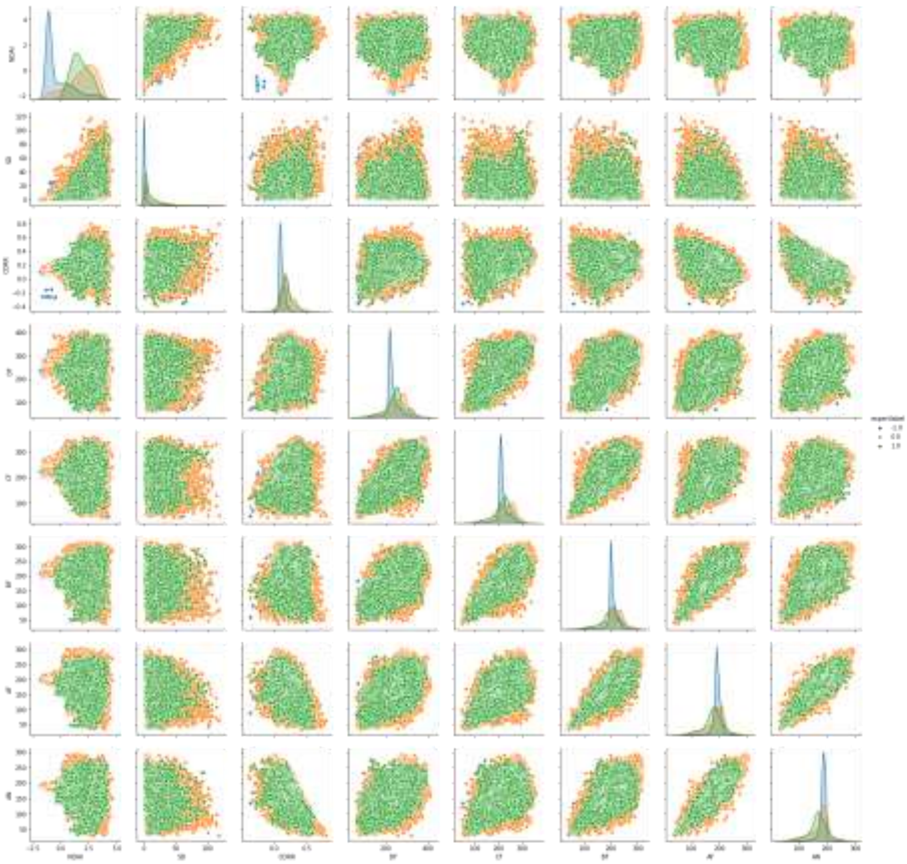


Image3

2a.We use two ways to split the data. In both ways, we use pixels in the third image as test set, because using a whole picture for testing can better simulate the real condition compared to using random sampling pixels from three pictures. Also, considering that the data is not i.i.d. in one picture, to avoid underestimate the error, it is better to use a whole picture as test set instead of using random sampling pixels from three pictures.

Furthermore, in the first way of splitting data, we use a quadrant split to generate validation set from the second image, because in this way the validation set looks like a miniature of the whole picture (this shape is similar to the whole image) and the split way can better measure the accuracy of the classifier methods.

In the second way of splitting data, we random sample 25 percent pixels from the second picture as the validation set. Although the data is not i.i.d., we can still use random sampling as long as using this way to split data will not influence accuracy of prediction.

2b.For the first splitting way:

The accuracy of a trivial classifier on the validation set: 0.6321

The accuracy of a trivial classifier on the test set: 0.2929

For the second way:

The accuracy of a trivial classifier on the validation set: 0.3755.

The accuracy of a trivial classifier on the test set: 0.2929

2c.We think three of the “best” features are NDAI, SD and CORR.

Our “best” feature criteria is the features perform very differently between clouded and not clouded pixels, so that we can use these features to give better classification between clouded and not clouded labels. We use quantitative and visual justification to choose three “best” features.First, we see mean, skewness and 95% range of the distribution of features grouped by labels.

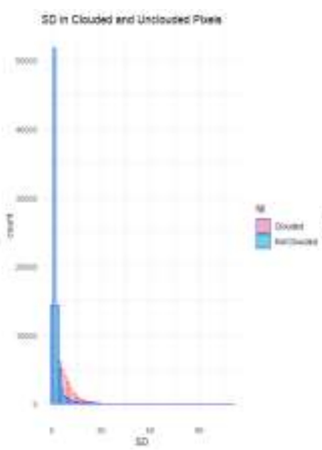
For the first splitting way, we get:

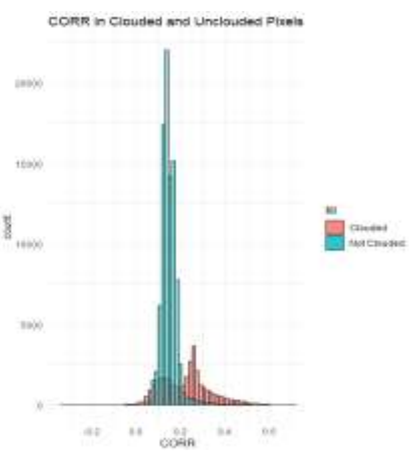
	expertlabel	mean_NDAI	mean_SD	mean_CORR	mean_DF	mean_CF	mean_BF	mean_AF	mean_AN
1	-1	-0.292730	2.797318	0.1486812	284.9524	270.7389	253.1984	229.2435	212.1399
2	0	1.715405	9.438050	0.1784599	282.9819	258.4193	236.5846	213.5470	198.6428
3	1	1.989370	8.157376	0.2242137	270.8700	238.1884	210.8494	187.1278	175.4867

	expertlabel	skewness_NDAI	skewness_SD	skewness_CORR	skewness_DF	skewness_CF	skewness_BF	skewness_AF	skewness_AN
1	-1	1.02593207	0.7521207	0.4507806	-0.5675108	-0.6791461	-0.7256161	-0.8412378	-0.6114869
2	0	-0.62523864	0.9606302	0.5386280	-0.5728911	-0.8952372	-0.9659967	-0.9753774	-0.9933609
3	1	-0.03234277	0.7840846	-0.2450420	-0.4523041	-0.7565783	-0.9953577	-1.0533966	-1.0826036

	expertlabel	range_1_NDAI	range_1_SD	range_1_CORR	range_1_DF	range_1_CF	range_1_BF	range_1_AF	range_1_AN
1	-1	-2.1906814	-6.796750	0.058529226	229.3206	212.7424	196.2320	176.3868	163.69035
2	0	-0.7372639	-8.857833	0.002446294	190.8151	169.5224	149.0961	130.0632	121.14190
3	1	0.7259287	-3.256677	0.001123264	160.3504	138.4887	115.1470	93.6047	84.40978

	expertlabel	range_2_NDAI	range_2_SD	range_2_CORR	range_2_DF	range_2_CF	range_2_BF	range_2_AF	range_2_AN
1	-1	1.605221	12.39139	0.2388331	340.5842	328.7355	310.1647	282.1002	260.5894
2	0	4.168074	27.73393	0.3544735	375.1487	347.3163	324.0731	297.0308	276.1437
3	1	3.252810	19.57143	0.4473041	381.3895	337.8880	306.5519	280.6510	266.5636





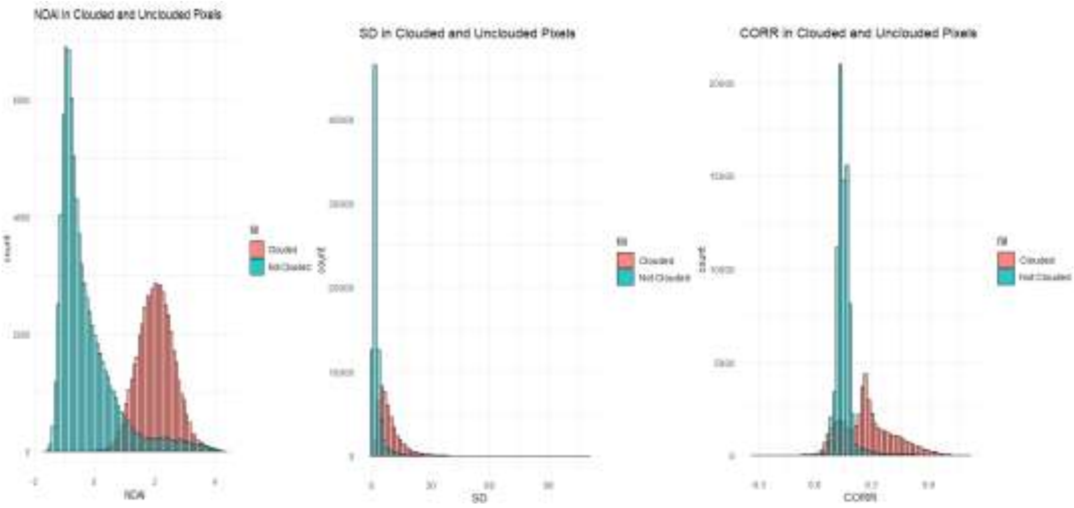
For the second splitting way, we get:

	expertlabel	..1	..2	..3	..4	..5	..6	..7	..8
1	-1	-0.2849006	2.756185	0.1486297	285.6195	271.2110	253.8850	229.9667	212.5852
2	0	1.7296699	9.447649	0.1788179	282.3904	257.5514	235.6725	212.7904	197.9878
3	1	2.0197317	9.319027	0.2747578	276.0093	235.1342	201.2381	173.9083	161.9568

	expertlabel	..1	..2	..3	..4	..5	..6	..7	..8
1	-1	1.007978506	0.7442283	0.3796017	-0.5518197	-0.6511695	-0.7002072	-0.8019212	-0.5791240
2	0	-0.606011764	0.9646181	0.5340383	-0.6015661	-0.9074937	-0.9655716	-0.9835736	-1.0037037
3	1	-0.002619026	0.8238853	0.2590212	-0.5494877	-0.5393932	-0.7008152	-0.8541549	-0.9164243

	expertlabel	..1	..2	..3	..4	..5	..6	..7	..8
1	-1	-2.1713420	-6.687070	0.0569165234	230.9983	214.0719	197.7050	177.91394	164.97139
2	0	-0.6975400	-8.873144	0.0003873568	189.3059	167.6487	147.1594	128.30808	119.69148
3	1	0.8473246	-5.016775	0.0003306323	172.5645	140.3611	104.5528	74.29846	63.86513

	expertlabel	..1	..2	..3	..4	..5	..6	..7	..8
1	-1	1.601541	12.19944	0.2403429	340.2408	328.3501	310.0650	282.0194	260.1990
2	0	4.156880	27.76844	0.3572484	375.4748	347.4542	324.1857	297.2728	276.2841
3	1	3.192139	23.65483	0.5491850	379.4541	329.9073	297.9234	273.5182	260.0484



The sign of skewness of NDAI in the cloud and not cloud label is different, as well as the CORR, and mean and range of two features in the two group are significantly different which can be seen in the histogram and numerical value above, so we choose these two features as "best" features. Besides, given that SD is the standard derivation of pixel values, different significantly from different labels, so we choose the feature as one of the three best features.

3a.Logistic:

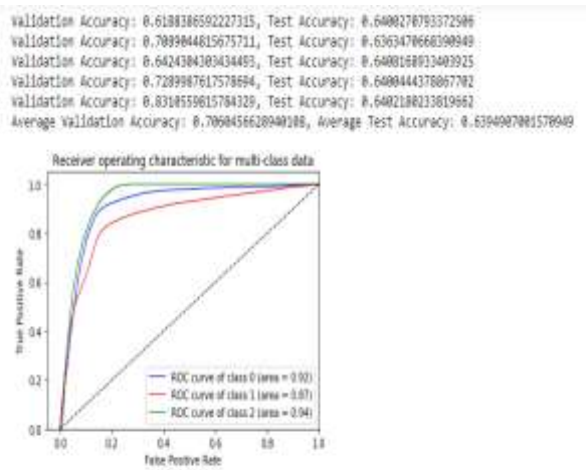
Assumption 1: binary logistic regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal.

Assumption 2, logistic regression requires the observations to be independent of each other. In other words, the observations should not come from repeated measurements or matched data.

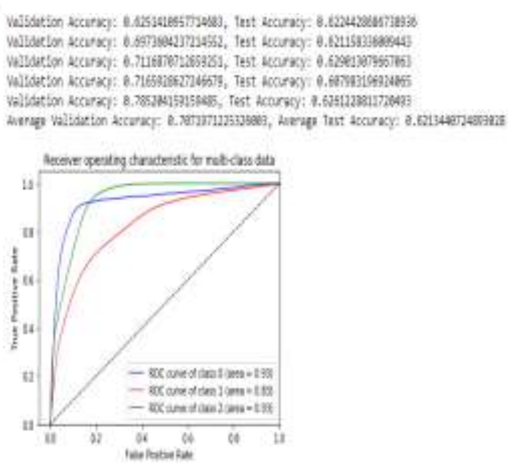
Assumption 3, logistic regression requires there to be little or no multi-collinearity among the independent variables. This means that the independent variables should not be too highly correlated with each other.

Assumption 4, logistic regression assumes linearity of independent variables and log odds. Although this analysis does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds.

Assumption 5, logistic regression typically requires a large sample size.   A general guideline is that you need at minimum of 10 cases with the least frequent outcome for each independent variable in your model.
In this case, assumption 1&3&5 are satisfied, but assumption 2 is not satisfied because pixels in one picture are not independent.
QDA: Assumption: The data is Gaussian, that each variable is shaped like a bell curve when plotted. In this case, we use qqplot to see if the data is Gaussian.
As we can see below, each feature is nearly Gaussian distribution.
Decision Tree: Assumption 1: The data can be described by features. Sometimes we assume these features are discrete, but we can also use decision trees when the features are continuous. Binary decisions are made on the basis of continuous features by determining a threshold that divides the range of values into intervals correlated with decisions.
Assumption 2: The class label can be predicted using a logical set of decisions that can be summarized by the decision tree.
Assumption 3: The greedy procedure will be effective on the data that we are given, where effectiveness is achieved by finding a small tree with low error.
In this case, all these three assumptions are satisfied.
KNN: Assumption: k-nearest neighbors assumes f(x) is well approximated by a locally constant function.
In this case, the assumption is satisfied.
For the two splitting ways, the test accuracy for the first way is slightly better than the second one generally. For the two splitting way, the logistic classification gives the best prediction, i.e. highest test accuracy(0.639,0.621), while the decision tree gives the worst prediction, i.e. the lowest(0.562,0.544)
3b.

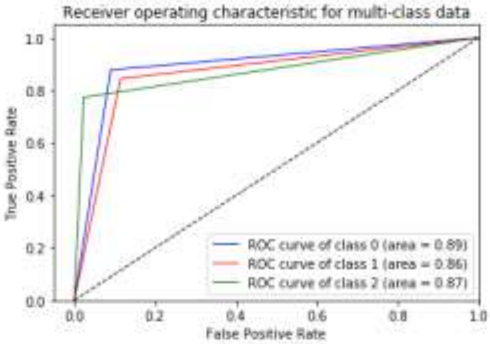


ROC logistic c1



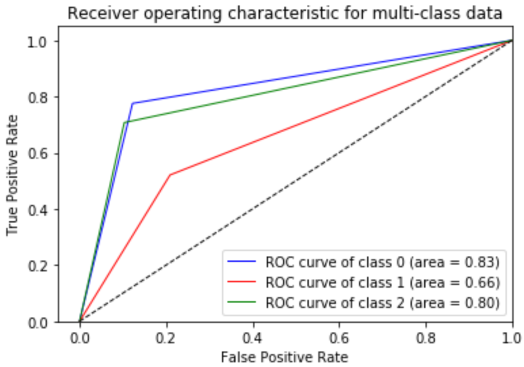
ROC logistic c2

```
Validation Accuracy: 0.5274911476549432, Test Accuracy: 0.5714868465591015
Validation Accuracy: 0.618534529576608, Test Accuracy: 0.5624083251603496
Validation Accuracy: 0.6146894619077619, Test Accuracy: 0.5507694177074564
Validation Accuracy: 0.7904763973671062, Test Accuracy: 0.562434362984629
Validation Accuracy: 0.8646623074750722, Test Accuracy: 0.5599954867771249
Average Validation Accuracy: 0.6831707687962982, Average Test Accuracy: 0.5614188878377323
```



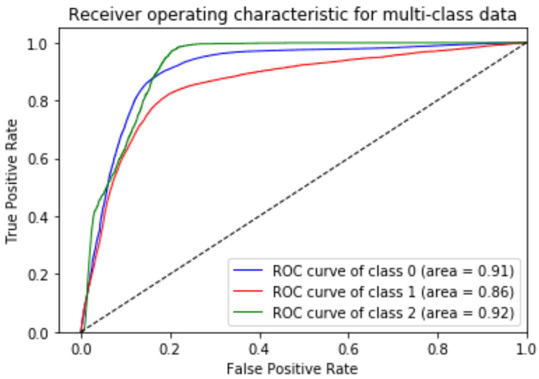
ROC Decision Tree C1

```
Validation Accuracy: 0.5314752105583052, Test Accuracy: 0.560689828757909
Validation Accuracy: 0.6161109663974993, Test Accuracy: 0.5570879297325916
Validation Accuracy: 0.6225796648432752, Test Accuracy: 0.5561418887837732
Validation Accuracy: 0.6745897369106538, Test Accuracy: 0.5471414808578595
Validation Accuracy: 0.7037141554692079, Test Accuracy: 0.5505263980141819
Average Validation Accuracy: 0.6296939468357883, Average Test Accuracy: 0.554317505229263
```



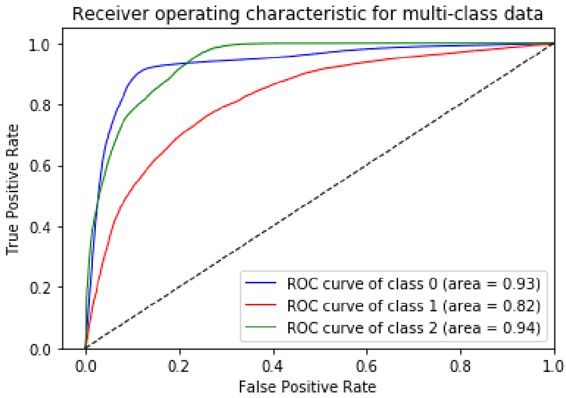
ROC Decision Tree C2

```
Validation Accuracy: 0.645645515173897, Test Accuracy: 0.6249338205299565
Validation Accuracy: 0.6850737514391849, Test Accuracy: 0.617061718322817
Validation Accuracy: 0.6730823539634609, Test Accuracy: 0.6040080890840761
Validation Accuracy: 0.7463558751330567, Test Accuracy: 0.6139024623102494
Validation Accuracy: 0.8244954706406273, Test Accuracy: 0.6003367558606802
Validation Accuracy: 0.7149305932700454, Average Test Accuracy: 0.6120485692215558
```



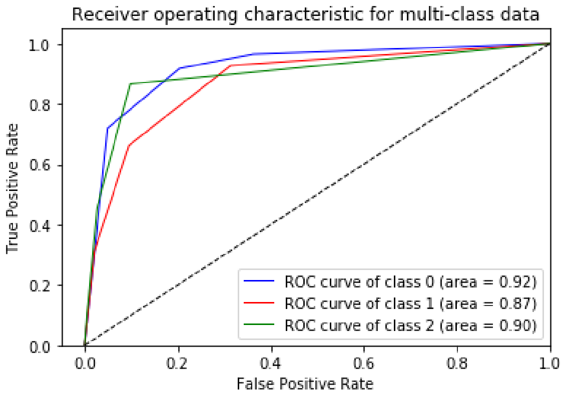
ROC QDA C1

```
Validation Accuracy: 0.6446991404011462, Test Accuracy: 0.611125094387113
Validation Accuracy: 0.6733307284883217, Test Accuracy: 0.6017080812727289
Validation Accuracy: 0.7148345923417556, Test Accuracy: 0.5990956195700288
Validation Accuracy: 0.7133368064600156, Test Accuracy: 0.5950076811581624
Validation Accuracy: 0.7937786267827295, Test Accuracy: 0.6106564135500837
Validation Accuracy: 0.7079959788947937, Average Test Accuracy: 0.6035185779876233
```



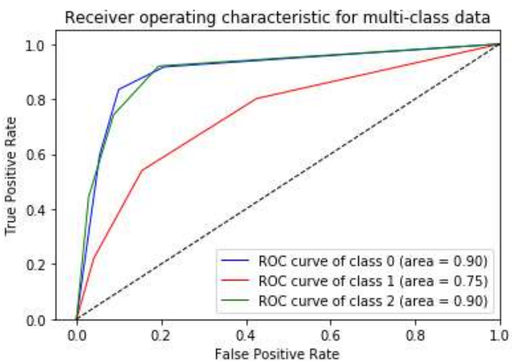
ROC QDA C2

```
Validation Accuracy: 0.5481067929528817, Test Accuracy: 0.6008835501705477
Validation Accuracy: 0.647839593335216, Test Accuracy: 0.590051815270316
Validation Accuracy: 0.6346316772749984, Test Accuracy: 0.5842974561045678
Validation Accuracy: 0.7480503117328873, Test Accuracy: 0.5941223951326627
Validation Accuracy: 0.8349444963395825, Test Accuracy: 0.5936016386470746
Validation Accuracy: 0.6827145743271131, Average Test Accuracy: 0.5925913710650338
```



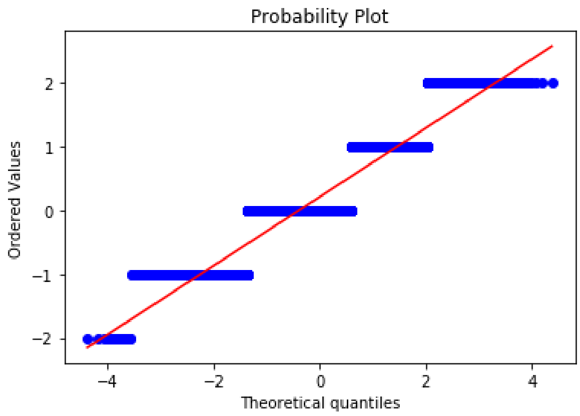
ROC KNN C1

```
Validation Accuracy: 0.7524475220873944, Test Accuracy: 0.5856601022418567
Validation Accuracy: 0.667203926576027, Average Test Accuracy: 0.5827733754567468
```

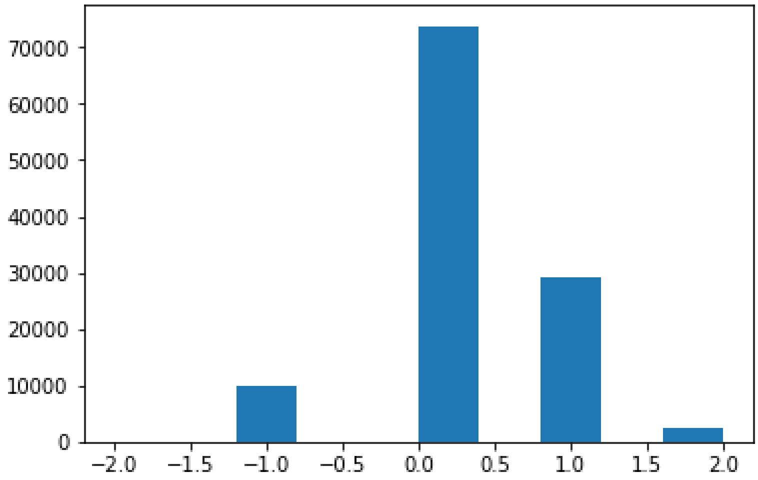


ROC KNN C2

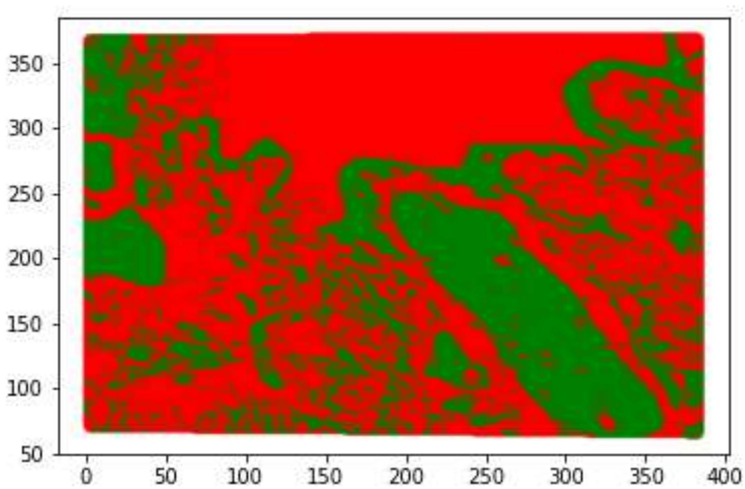
4a



From the plot we can see that it's basically normal distribution for qq plot but there are some patterns for the error. This QQ plot shows basically a normal pattern for the residual so the model is a good fit and kurtosis of the residual is basically normal



From the histogram we can see that the majority of the labelings are correct with the second highest the mistake (-1 and 1) due to the unlabeled area. So in general this classification method is good but if the unlabled prediction can be improved then the
#result can be significantly improved.
4b



Red: correct; green: mistake (classification2)

Yes there are some patterns. The green area represents the mistaken area. In this case, those coordinates in the upper left and

lower right are most likely to be mistaken. Although the majority of mid part is correct, there's still some mistaken area inside the middle.

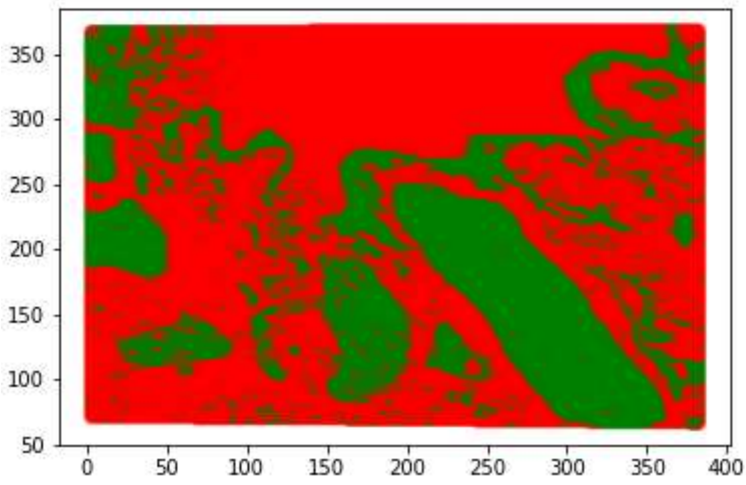
```
mt = g2_test[g2_test["mistake"] == 0]
pd.DataFrame.describe(mt)
```

	x	y	expertlabel	NDAI	SD	CORR	DF	CF	BF	AF	
count	43077.000000	43077.000000	43077.000000	43077.000000	43077.000000	43077.000000	43077.000000	43077.000000	43077.000000	43077.000000	430
mean	203.271792	197.031618	0.227221	1.418255	9.617976	0.192494	248.471196	223.150351	202.540191	182.687063	1
std	114.382645	78.794348	0.725674	1.257315	11.300348	0.098055	43.921705	38.995299	36.313790	34.283500	
min	2.000000	65.000000	-1.000000	-1.590378	0.275784	-0.361673	64.704857	47.871189	40.983593	33.662514	
25%	106.000000	120.000000	0.000000	0.539731	2.586078	0.130923	223.671020	204.910750	187.323300	169.092850	1
50%	224.000000	194.000000	0.000000	1.393026	5.028268	0.173935	249.574140	223.565050	204.694000	189.095760	1
75%	302.000000	258.000000	1.000000	2.448245	12.288973	0.225657	276.785690	249.998000	228.126270	205.645265	1
max	383.000000	369.000000	1.000000	4.379109	103.559840	0.750892	406.454770	358.046540	315.545750	300.467680	2

	x	y	expertlabel	NDAI	SD	CORR	DF	CF	BF	
count	115217.000000	115217.000000	115217.000000	115217.000000	115217.000000	115217.000000	115217.000000	115217.000000	115217.000000	115217.0C
mean	193.144675	218.328302	-0.108560	1.267211	10.399983	0.167762	246.507080	222.810895	204.318173	186.63
std	110.263796	87.074966	0.682308	1.511791	12.921442	0.090818	44.772517	39.879922	36.516583	33.75
min	2.000000	65.000000	-1.000000	-1.841971	0.198708	-0.387243	61.032272	43.656754	37.539799	33.66
25%	98.000000	143.000000	-1.000000	0.018797	1.566069	0.105517	221.368730	206.536320	192.686480	175.17
50%	193.000000	218.000000	0.000000	1.472049	5.058567	0.155584	243.449970	221.233860	205.358690	191.78
75%	289.000000	294.000000	0.000000	2.514264	14.339034	0.206516	277.189540	250.335160	227.550700	205.81
max	383.000000	369.000000	1.000000	4.563939	117.581020	0.789216	410.527100	360.683560	315.545750	304.06

If we compare the two tables, we can see that NDAI, SD and CORR's mistakes lie within the range of the whole test data set.

4c. One of the better methods is to include clustering method and this for sure will improve because of the clustering patterns we discovered before because the clustering can make more unlabeled data labeled which will at least not decrease the accuracy(and very likely to improve accuracy)



Classification1

	x	y	expertlabel	NDAI	SD	CORR	DF	CF	BF	AF	
count	41453.000000	41453.000000	41453.000000	41453.000000	41453.000000	41453.000000	41453.000000	41453.000000	41453.000000	41453.000000	414
mean	205.056449	199.131667	0.314718	1.316444	8.777940	0.184381	249.622928	224.078397	203.981028	184.870688	1
std	110.965351	76.992320	0.758435	1.207128	10.321415	0.088505	43.226927	37.510067	33.796199	31.374218	
min	2.000000	65.000000	-1.000000	-1.590378	0.275784	-0.361673	64.704857	47.871189	48.319122	38.565712	
25%	117.000000	133.000000	0.000000	0.504636	2.506477	0.129083	224.314010	205.914990	189.435330	171.922550	1
50%	224.000000	194.000000	0.000000	1.302842	4.780944	0.171237	250.154540	224.381450	205.467060	189.505170	1
75%	300.000000	259.000000	1.000000	2.223672	11.057730	0.215628	277.548370	250.556720	227.326740	205.580250	1
max	383.000000	369.000000	1.000000	4.379109	103.559840	0.729798	391.251500	348.389470	315.545750	300.467680	2

	x	y	expertlabel	NDAI	SD	CORR	DF	CF	BF	
count	115217.000000	115217.000000	115217.000000	115217.000000	115217.000000	115217.000000	115217.000000	115217.000000	115217.000000	115217.00
mean	193.144675	218.328302	-0.108560	1.267211	10.399983	0.167762	246.507080	222.810895	204.318173	186.63
std	110.263796	87.074956	0.682308	1.511791	12.921442	0.090818	44.772517	39.879922	36.516583	33.79
min	2.000000	65.000000	-1.000000	-1.841971	0.198708	-0.387243	61.032272	43.856754	37.539799	33.66
25%	98.000000	143.000000	-1.000000	0.018797	1.566069	0.105517	221.368730	206.536320	192.686480	175.17
50%	193.000000	218.000000	0.000000	1.472049	5.058567	0.155584	243.449970	221.233860	205.358690	191.78
75%	289.000000	294.000000	0.000000	2.514264	14.339034	0.206516	277.189540	250.335160	227.550700	205.81
max	383.000000	369.000000	1.000000	4.563939	117.581020	0.789216	410.527100	360.683560	315.545750	304.06

The plot of the two classifications show almost the same pattern as well as the table does. Thus changing classification doesn't affect the result.

4e

Thus, we conclude that the logistic regression has the best performance with classification 1 slighter better than classification 2.

Without expert labels, using the logistic classification leads to some mistake in labeling and the mistake clusters in the upper left and lower right of the whole. The range of the features for the mistaken ones lies within the range of the whole test data set for the three main features. Changing the classification does not change the result significantly. Due to the clustering patterns we discovered, a good approach might be to introduce clustering in the future which will consider the extent of cloudiness (unsupervised learning) that will improve the result.

Github website: https://github.com/cosmos139/Stats-154.git