Lab 4 - Cloud Data, Stat 215A, Fall 2019

Aya Amanmyradova
Spencer Wilson
Ziyang Zhou

November 21, 2019

1 Introduction

As global warming becomes more and more a reality, global climate models predict that surface air temperatures and atmospheric carbon dioxide levels will increase throughout this century. The Arctic is one of the regions where global warming has the most impact. The change in distribution of ice covered surfaces and clouds can further accelerate global warming. Being able to collect accurate data from satellites would immensely help to study cloud coverage. This became possible by the launch of Multiangle Imaging SpectroRadiometer (MISR) onboard the NASA Terra satellite in 1999, which takes radiation measurements at 9 view angles. Therefore, the goal of this report is to model cloud detection based on measurements obtained from MISR. In order to achieve this, after proper exploratory data analysis we develop several classifiers, asses their fit using several metrics and choose the best classification model. For this study, we used logistic regression, random forest, quadratic discriminant analysis and naive bayes.

2 Exploratory Data Analysis

2.1 The Data

The dataset consists of 3 images from the satellite, seen in Figure 1. For every pixel, we have x and y coordinates and expert labels (cloud = +1, not cloud = -1, unlabeled = 0), along with 8 other variables: NDAI, SD, CORR, DF, CF, BF, AF and AN. The last 5 variables are radiances obtained from cameras located at different angles. The first 3 measures are features derived from radiances to differentiate surface pixels from cloudy ones. NDAI is a normalized difference angular index that compares mean radiation collected from DF (zenith angle) and AN (nadir direction) cameras. SD is a standard deviation within groups of nadir camera radiation measurements, and CORR is an average linear correlation of radiation measurements at different view angles.

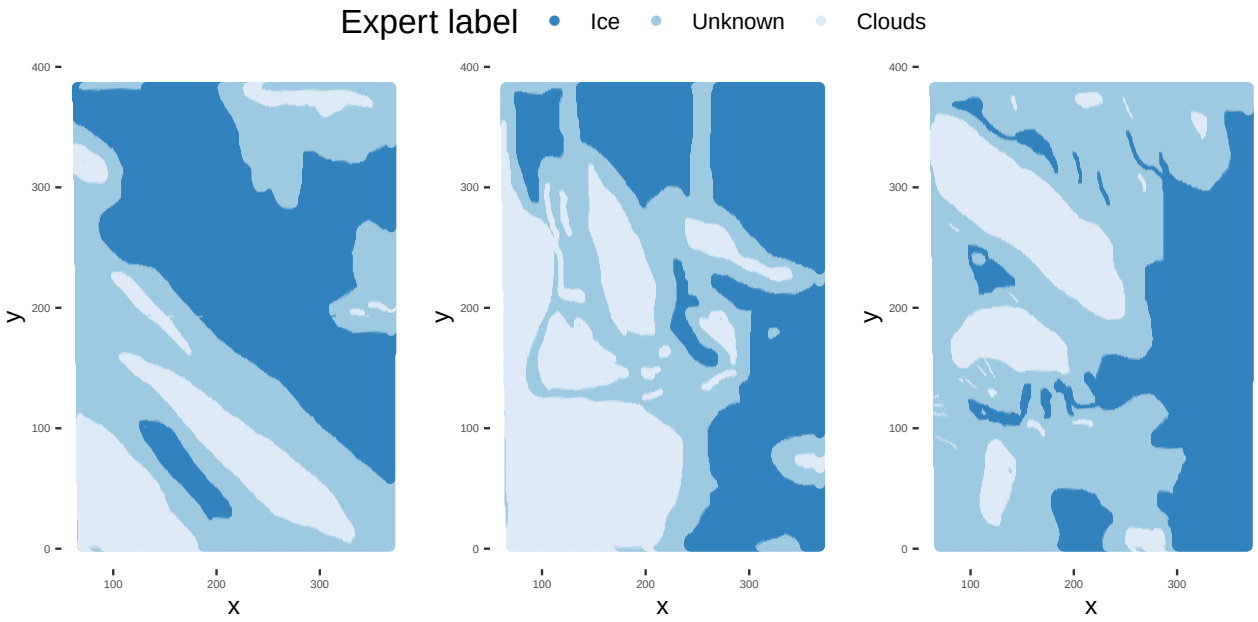


Figure 1: Expert labels for the presence or absence of clouds, according to a map.

As we can see from Figure 2, image 2 has the most 1-label, while image 3 has the most 0-label (unlabeled), meaning that it contains the most unwanted data. For each image, there are no missing values. The typical IID assumption is broken in this case because the cloud appears in patches. This means that if one pixel is clssified as cloudy, then the surrounding pixels are also highly likely to be cloudy. This geographic property causes correlation structure among pixels, so we would need to properly compensate for this violation of IID assumtpion in the processing stage.

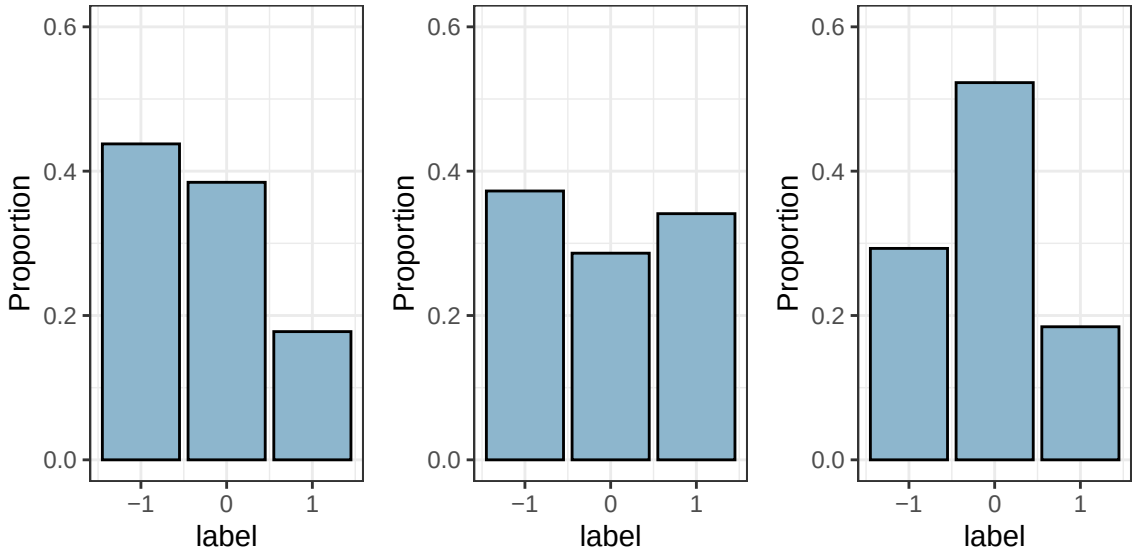


Figure 2: Proportion of Labels in Each Image

2.2 The Relationship Between Variables

We explored the relationships between the variables both visually and quantitatively. The pair-wise relationships between features is shown in Figure 3. We see that the five angular features, most certainly, are highly correlated with one anohter. However, the correlations were even stronger for non-cloud pixels than cloud pixels. The average correlation between all pair-wise radiances for non-cloud pixels was 0.94, while the average correlation for the cloud pixels was 0.80. Other features are moderately related to each other in either a positive or a negative way but there is no clear pattern of any relationship. The average correlation between NDAI, SD and CORR for cloud pixels was 0.53 and for non-cloud ones was 0.62. To further understand the features, we made a overlaying label distribution plot for each feature in Figure 4. We observe that the label distributions are very similar among AF, AN, BF, CF but quite divergent in CORR, log(SD), and NDAI. This is not surpursing since CORR, SD, and NDAI are the features specifically created for this classification task, as mentioned in the original paper Yu (2008). We performed a log-transformation on SD since its original distribution is highly skewed to the right. Most of the distributions here are bell-shaped. Specifically for NDAI, we observe that the distributions for 1-label and -1-label almost have no overlap, thus this suggests NDAI could be a good predictive feature in this task. These findings suggest that NDAI, SD and CORR might be good predictors of presense of clouds, while radiance measures are highly correlated and using only one of them is sufficient for prediction.

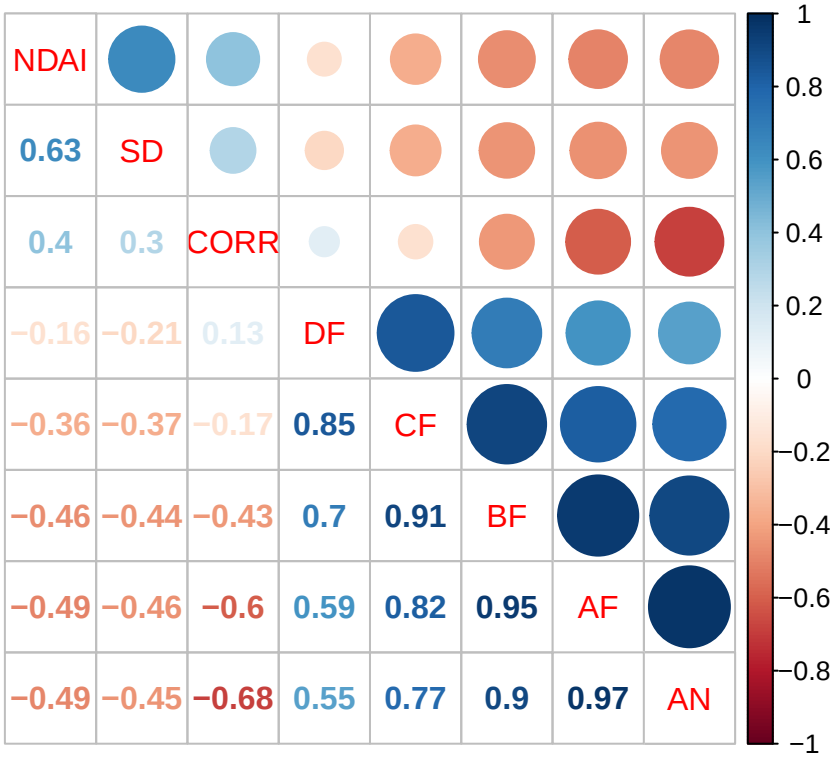


Figure 3: Correlation Plot among Potential Features

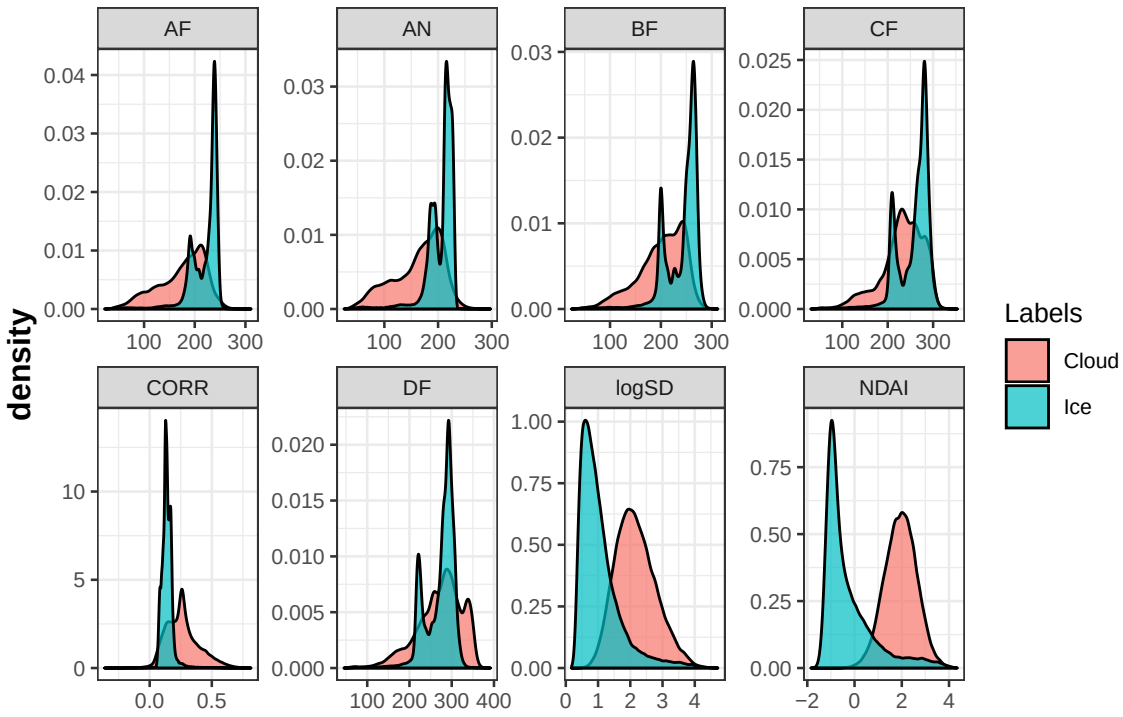


Figure 4: Overlaying Distribution of Labels among Features

3 Modeling

3.1 Data Splitting Method

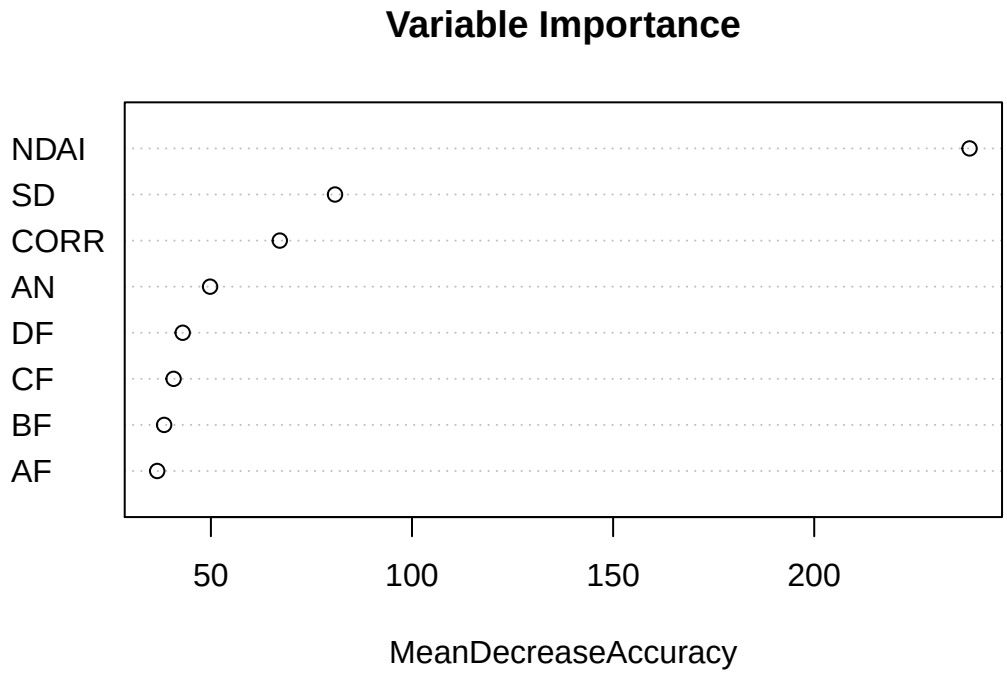
In order to compensate for violation of IID assumption, we suggest other methods of splitting data instead of doing a trivial 80-20 splitting on the entire data set. The first method is to sample based on label, i.e., 80-20 stratified sampling on labels for each image. The reason of doing so is because as Figure 4 shows, the images are quite different and simply doing a 80-20 splitting on the entire data set would result in unbalanced labels among images. For example, it would be possible that less number of pixels with cloud in image 3 will be sampled becasue of unbalanced labels. The second method we suggest is block sampling, i.e., we divide the image into grids according to (x, y) coordinates and do a 80-20 splitting within each block. Through this method, we would be able to ensure that small patches of clouds will also be captured in training data. In later analysis, we will employ both method and apply those on cross validation. Another possible way of splitting the data accounting their correlation structure is to find the center of clouds, cirle the contours, find the boundaries of each cloud, and sample within/without the boundaries accordingly. Since this share similar characteristics as label splitting, we will thus only implement the first two.

3.2 Feature Selection

As we noted in our exploration of variables, the various radianace angles are highly correlated with one another. Consequently, it may not be necessary to include all five when constructing the classification model. We performed feature selection via random forests using a training test set containing all three images split by label. The three best features based on the mean decrease in accuracy are clear from the variable importance plot: NDAI, SD, and CORR. These results fit well with our expectations from inspecting the distribution of all features in Figure 4. NDAI is far and away the most salient feature, with SD and CORR trailing before all of the radiances. Still, the radiance angles all had a notable range with little overlap where the density of

non-cloud pixels spikes that a model could exploit. The final feature space included NDAI, SD, CORR, and the most importance angular feature: AN.

```
##             -1        1 MeanDecreaseAccuracy MeanDecreaseGini
## NDAI 291.16181 77.47064            238.61707        26416.997
## SD    66.12240 35.53348             80.85954        15712.920
## CORR  45.25342 94.36993             67.12971        12917.994
## DF    33.65753 59.60923             42.98817         4463.038
## CF    34.81647 45.38045             40.72497         3178.992
## BF    35.98886 21.35161             38.37628         4067.287
## AF    34.76503 14.33157             36.65152         5592.026
## AN    45.79725 24.14611             49.79958         6857.972
```



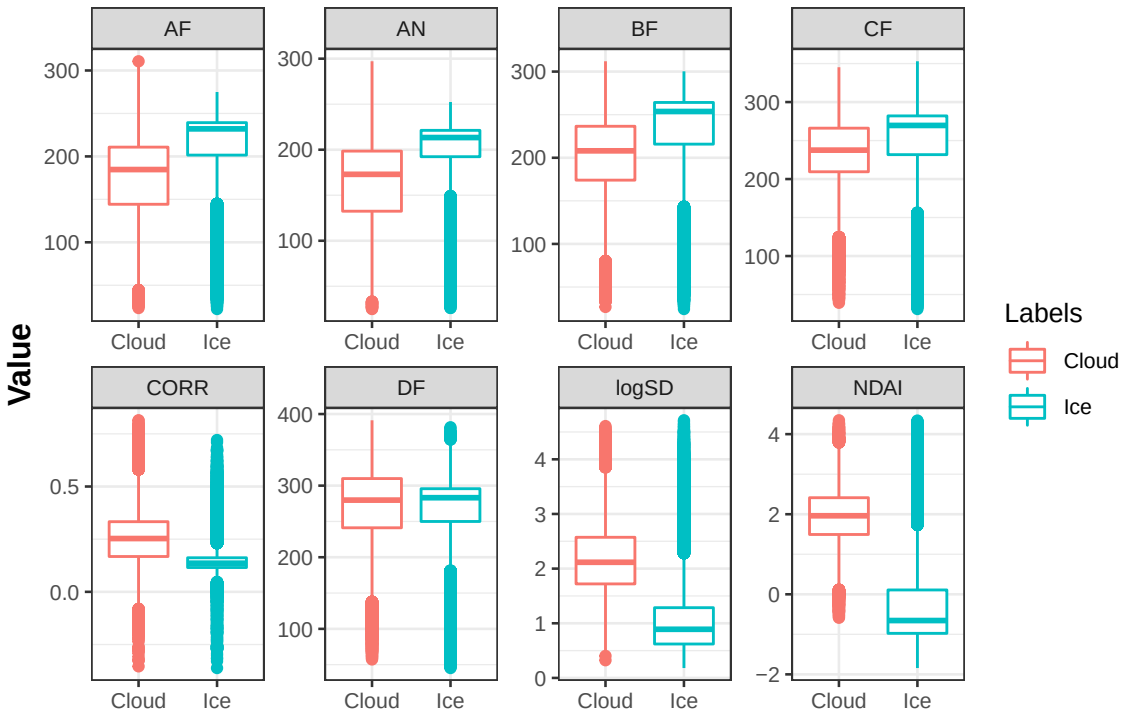


Figure 5: Boxplot Distribution of Labels among Features

3.3 Description and Assumptions of Classifiers

We developed four classification models to test the presence of clouds: logistic regression, random forest, quadratic discriminant analysis, and naive Bayes. We included 4 features: NDAI, SD, CORR and AN, the choice of which was justified above.

- **Logistic regression** models binary dependent variable. It assumes that there is no multicollinearity within features. It is obvious that this assumption is violated in the data, especially in regards to the different radiance angles. Therefore, we chose not to include all radiance measures. Since our aim is prediction and not inference we can be lenient with this assumption. It also assumes independence of error terms with features, and linearity of features with log odds.
- **Random forest** consists of a large number of decision trees that operate as an ensemble. Each tree in the random forest outputs a class prediction, and the class with most votes becomes the model's prediction. It does not assume any formal distribution and is non-parametric.
- **Quadratic discriminant analysis** is an extension of linear discriminant analysis, where the assumption of equal variance of classes is relaxed. However, it still assumes that features are drawn from multivariate Gaussian distribution and that $p < n$.
- **Naive Bayes** is a classifier based on Bayes' theorem. It has a strong asssumption of independence of features. We know this does not hold true in our data, and indeed, this classifier performed worst among all four.

3.4 Assessment of fit of Classifiers

Classifier fit was assessed using k-fold cross-validation and the F1-score for evaluation. As the harmonic mean of precision and recall, the F1-score balances identifying clouds while not diminishing the value of the classifier by over application. With k set to five, we compared all of our classification models using both

the label and block methods of training/ test splitting that were previously mentioned. Random forests are unmatched in their performance: their average score of 0.93 is 7% higher than the second best model. Still, there is no getting away from the fact that their structure necessitates a long amount of time to run. The other three predictors are much quicker, and among them QDA consistently outperforms both Logistic regression and Naive Bayes. If iteration or runtimes are critical, we would recommend QDA, but the best results are achieved with random forest. Results were consistent between both splitting techniques. Label splitting is faster and easier to implement than blocking, so is the one we recommend when using multiple datasets with uneven class distribution.

Table 1: Cross-Validation Comparison

Label/ Block	logistic	rf	qda	nb
CV1	0.8347833	0.9324412	0.8674893	0.8326855
CV2	0.8331583	0.9296997	0.8683517	0.8336917
CV3	0.8337691	0.9285498	0.8667568	0.8276057
CV4	0.8318349	0.9282007	0.8661101	0.8358463
CV5	0.8334741	0.9311468	0.8661375	0.8307680
Avg	0.8334039	0.9300076	0.8669691	0.8321194
CV1	0.8332254	0.9323440	0.8659395	0.8331304
CV2	0.8347104	0.9281462	0.8709640	0.8318735
CV3	0.8321051	0.9294991	0.8676550	0.8354615
CV4	0.8330034	0.9335969	0.8645817	0.8350179
CV5	0.8317407	0.9306391	0.8666203	0.8262129
Avg	0.8329570	0.9308451	0.8671521	0.8323393

3.5 Post-hoc EDA for random forest

To study patterns in misclassification errors of random forest, we plotted all three images showing misclassified pixels along with correct ones on a map (see Figure 6). As we can see, there are specific places where the classifier fails to classify pixels correctly. In image 1, at the bottom left majority of the non-cloud surface was misclassified, probably because it is in close proximity to clouds from both sides. The same misclassification occured in image 3. A small non-cloud surface was classified as cloud, since it is surrounded with cloud pixels from three sides. From this visual assessment, it seems that random forest classifies cloud pixels better than non-cloud pixels. However, we need to look at the quantitative evidence too. In order to do that, we calculated the percentage of misclassified units for each label. For images 1, 2 and 3, the percentage of misclassification for cloud pixels is 8.5, 0.8, and 10.7 respectively. Correspondingly, the percentage of misclassified pixels for non-cloud pixels is 4.7, 4.3, and 7.9 respectively. Across all images, the percentage of misclassification for cloud pixels is 5.36 and for non-cloud pixels is 5.41. The difference is minimal, therefore we do not believe that there is a difference in misclassification according to labels. In addition, the misclassification rate was lowest for image 2. This may be due to the low number of unlabeled pixels in this image. Furthermore, we visually checked for patterns in misclassification errors based on ranges of feature values (Figure 7). It appears, that the classifier tends to missclassify pixels with lower average radiance values from nadir camera and higher average NDAI values.

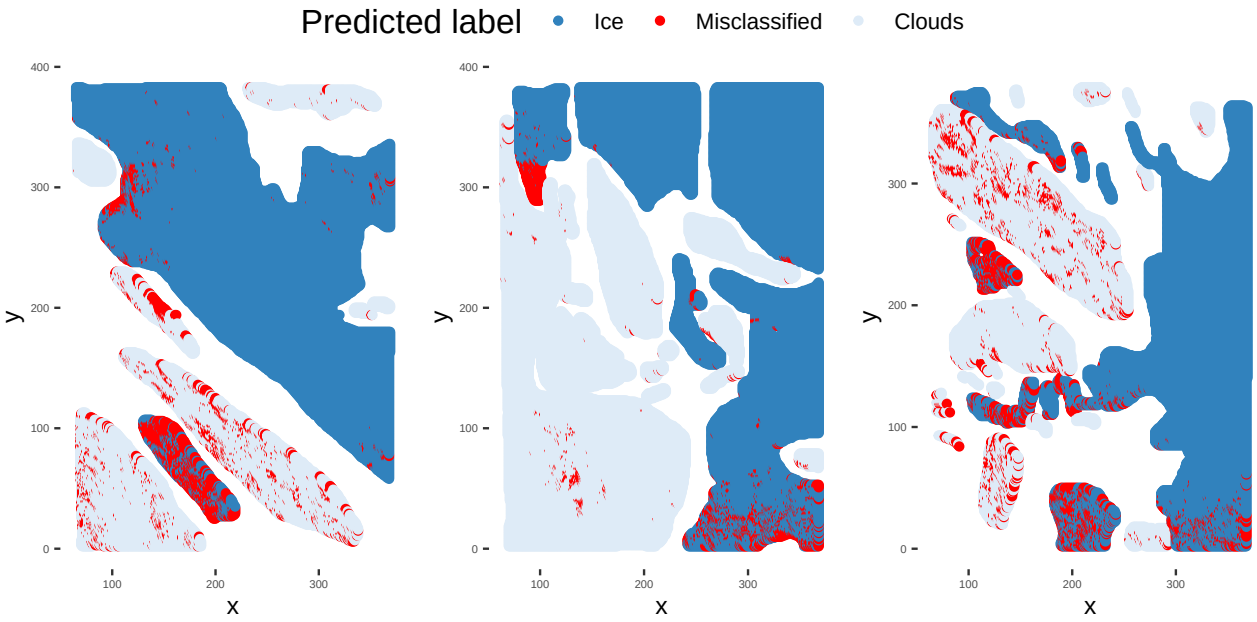


Figure 6: Misclassified labels

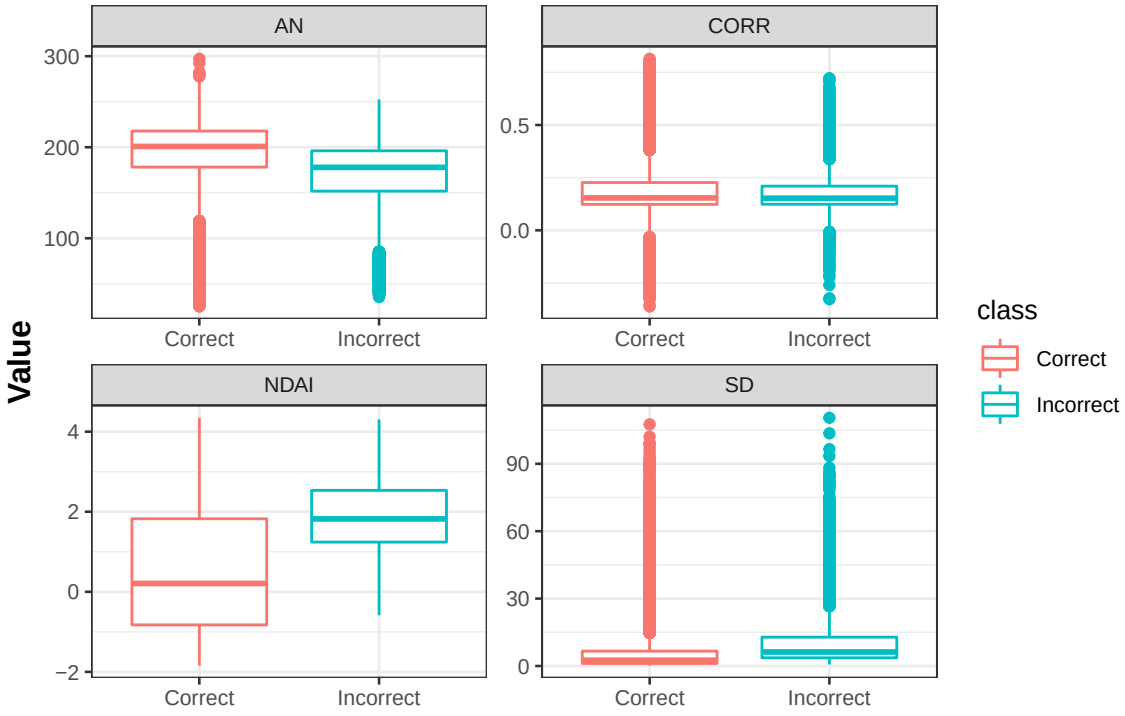


Figure 7: Misclassification based on feature values.

4 Conclusion

By comparing our models above, we observe an increase in classification metrics when model complexity increases. However, even though Random Forest turns out to have the highest score, it takes a much longer

amount of time for it to run because it needs to fit a large number of decision trees and output a classification result based on a voting mechanism. Therefore, there are some trade-offs between the model complexity and performance, if the running time is not a big issue, then Random Forest would be better, otherwise QDA is preferred since it is stable and fast. As we first attempt to try some classification models on the data set, we found out that fitting SVM and KNN are too computationally expensive. However, when we have more images in the feature, running time would be an important factor in choosing a better model. Also, more complex models also have difficulty in visualization. For example, we are able to plot a single decision tree, but are not be able to plot the entire forest in the training process of random forest, so in some way random forest serves as a black box where we don't have full information about what is going during training. Another factor of consideration is the model assumptions. Since we have cloud image data that clearly violates the IID assumption required by most of statistical models and we are not be able to standardize features since standardized angles don't make physical sense, models with less constraints would be preferred in our case. For example, Naive Bayes assume IID data, so we would not be so surprised by its bad performance on the task. With that in mind, we carefully dealt with the correlation structure in the data using both label splitting and block splitting so that we wouldn't miss a small patch of cloud that causes the problem of covariates shift. Moreover, if, in the future, we would like to apply our classification rule to more images, then it would also be very essential for us to get a good amount of images to learn. Lastly, in the diagnostic section, we showed that the feature importance output is consistent with our correlation analysis that CORR, SD, NDAI are generally more important in cloudiness classification. In conclusion, our analysis showed that effective statistical analysis could help natural science studies to be more efficient, and instead of labeling all the clouds by hands, we could just pass in the images into the trained model and let it make the decisions.