DISASTER RELIEF PROJECT

HUILIN CHANG
UNIVERSITY OF VIRGINIA, SYS6018

SUMMARY TABLE

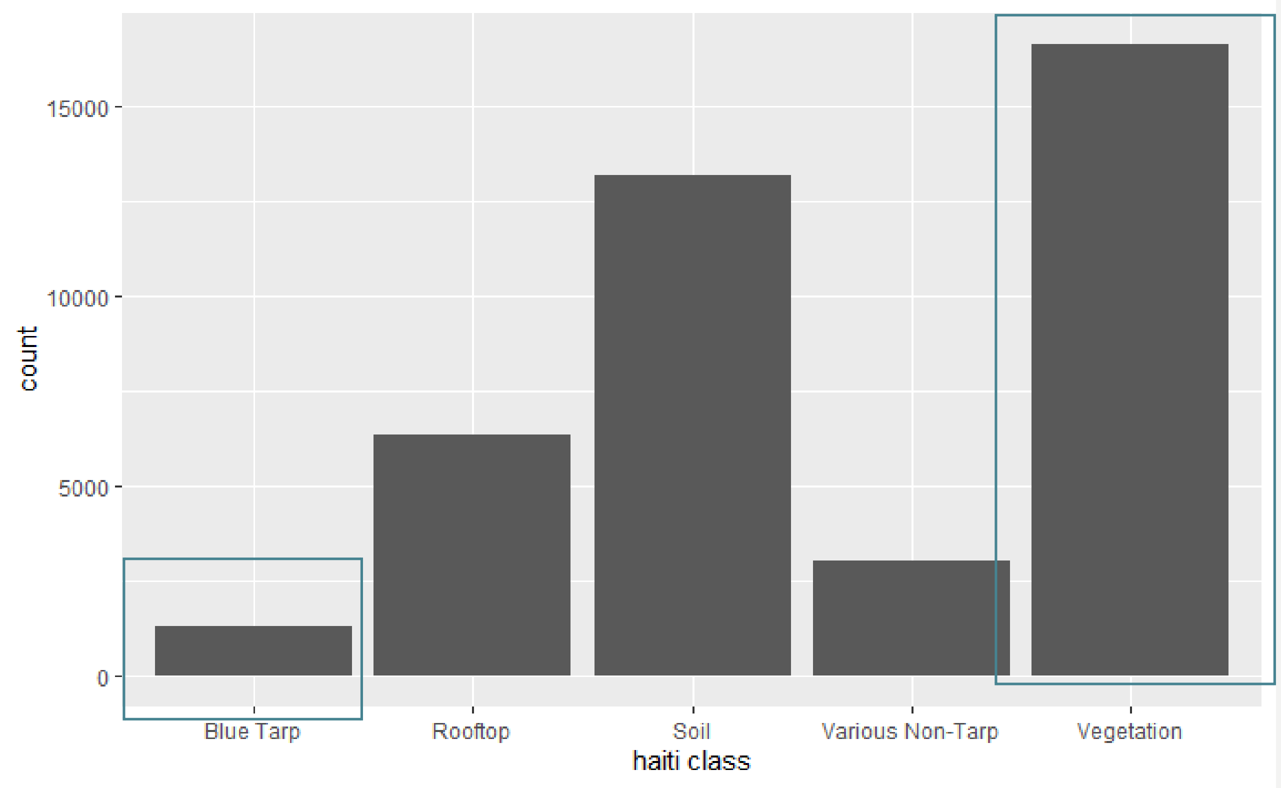
	Accuracy	Sensitivity (Recall)	Specificity	F measure	AUC
KNN(K=13)	92.89	96.54	99.72	95.23	99.85
LDA	85.05	80.45	99.98	88.80	99.35
QDA	86.88	86.88	99.87	91.05	99.55
Logistic regression	88.36	91.34	99.74	91.68	99.75

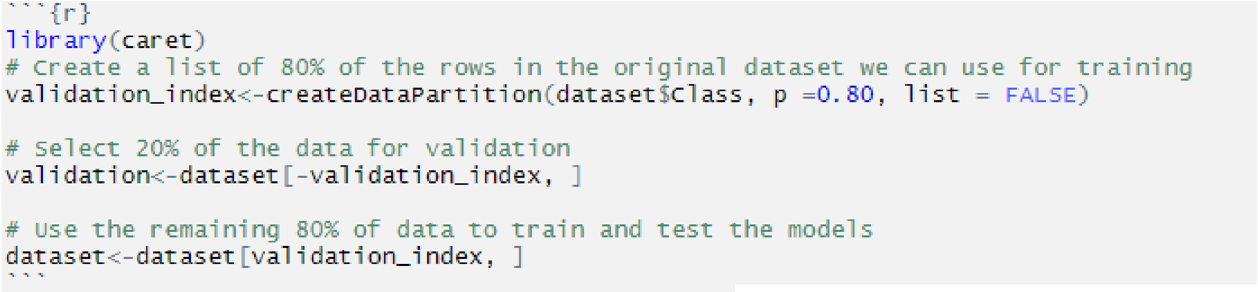
BACKGROUND

- A real historical data-mining problem, locating displaced persons living in makeshift shelters following the destruction of the earthquake in Haiti in 2010.
- people whose homes had been destroyed by the earthquake were creating temporary shelters using blue tarps.
- 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.

MODEL DEVELOPMENT

- **Nature of data**
  - dim(data) = 63241 * 4
  - Class = five
- **Model Considerations**
  - KNN
  - LDA
  - QDA
  - Logistic Regression
- **Two approaches**
  - Broaden the class because the concern is blue tarp specific
  - Keep five classes





~~~
```{r}
library(caret)
# Create a list of 80% of the rows in the original dataset we can use for training
validation_index<-createDataPartition(dataset$Class, p =0.80, list = FALSE)

# Select 20% of the data for validation
validation<-dataset[-validation_index, ]

# Use the remaining 80% of data to train and test the models
dataset<-dataset[validation_index, ]
```
~~~

METHODOLOGY

- Five classes:
  - Blue Tarp
  - Rooftop
  - Soil
  - Various Non-Tarp
  - Vegetation
- The distribution of the five classes is uneven, blue tarp is ~ 3.2 %
- Split the samples into 20%/80% ratios – validation/training sets
- Adopt 10 fold CV
- The scatter-matrix shows the attributes distribution of five classes

Five classes:

Blue Tarp

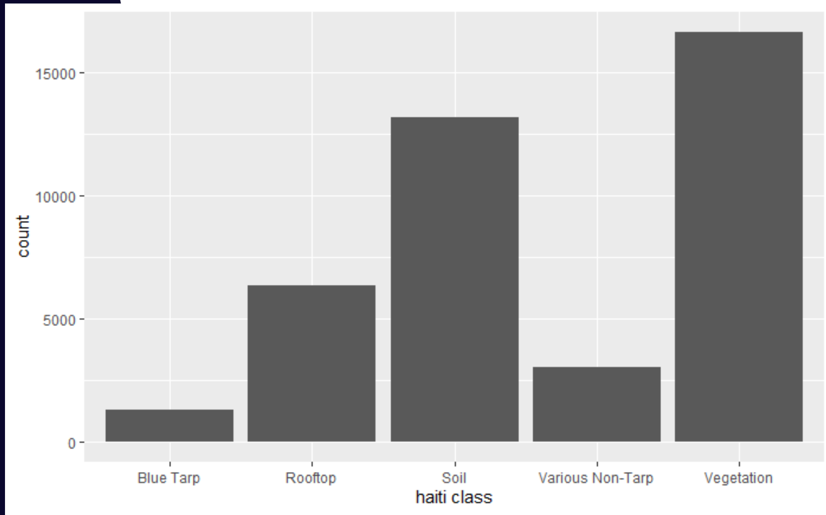
Rooftop

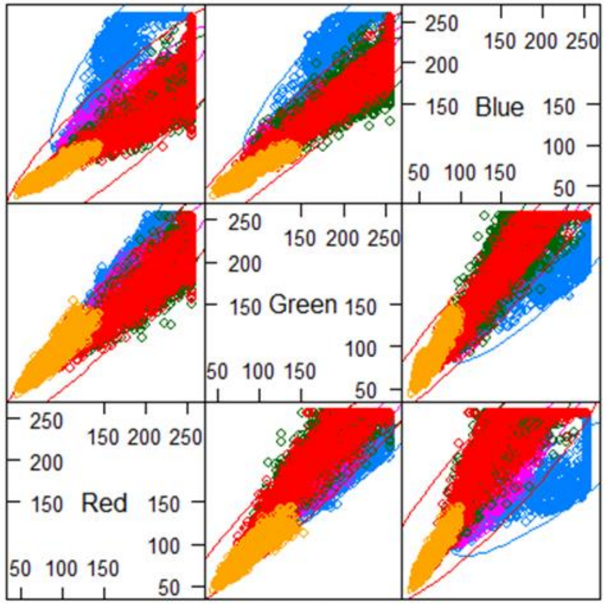
Soil

Various Non-Tarp

Vegetation

Our interest Blue Tarp is 3.20%

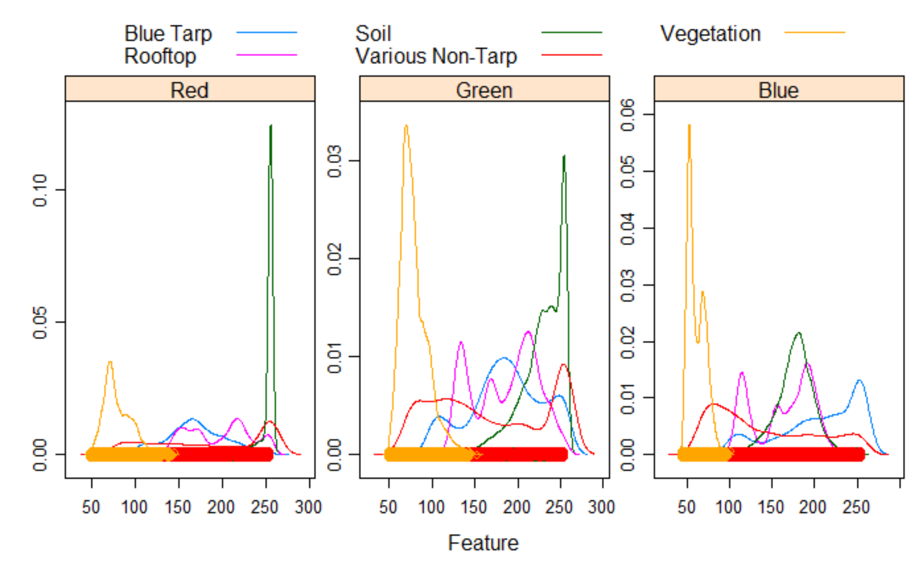


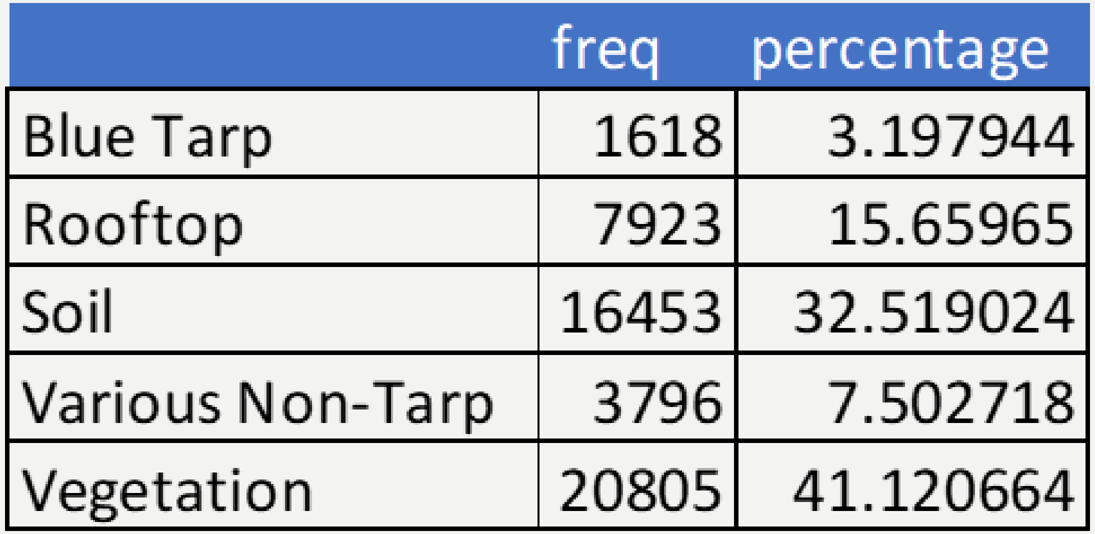


Scatter Plot Matrix

5 CLASSES ON FEATURES UNDERSTADNING

- Blue Tarp: 3.20%
- Rooftop: 15.66%
- Soil: 21.52%
- Various Non-Tarp: 7.50%
- Vegetation: 41.12%

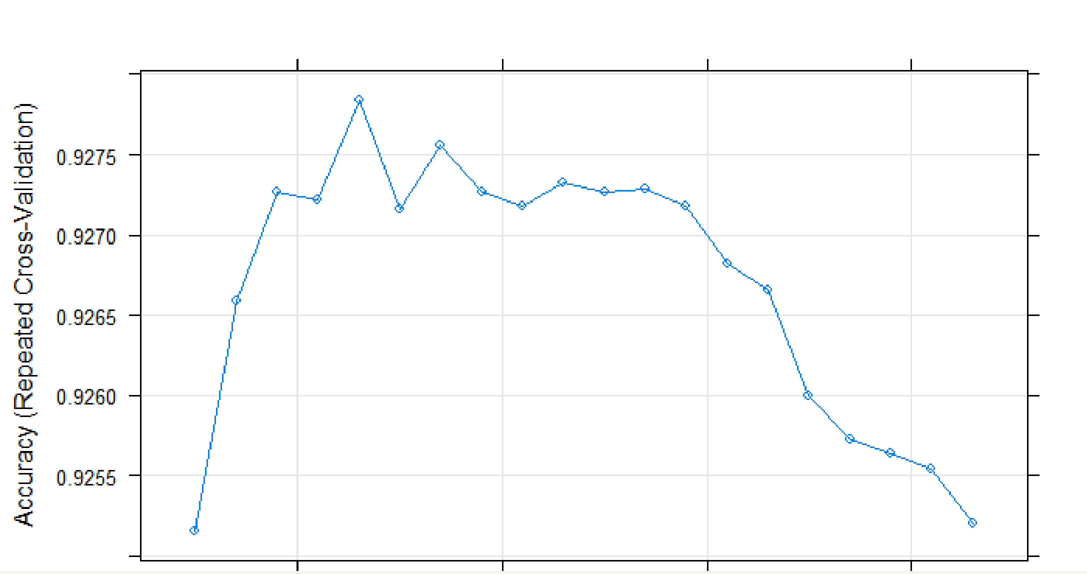


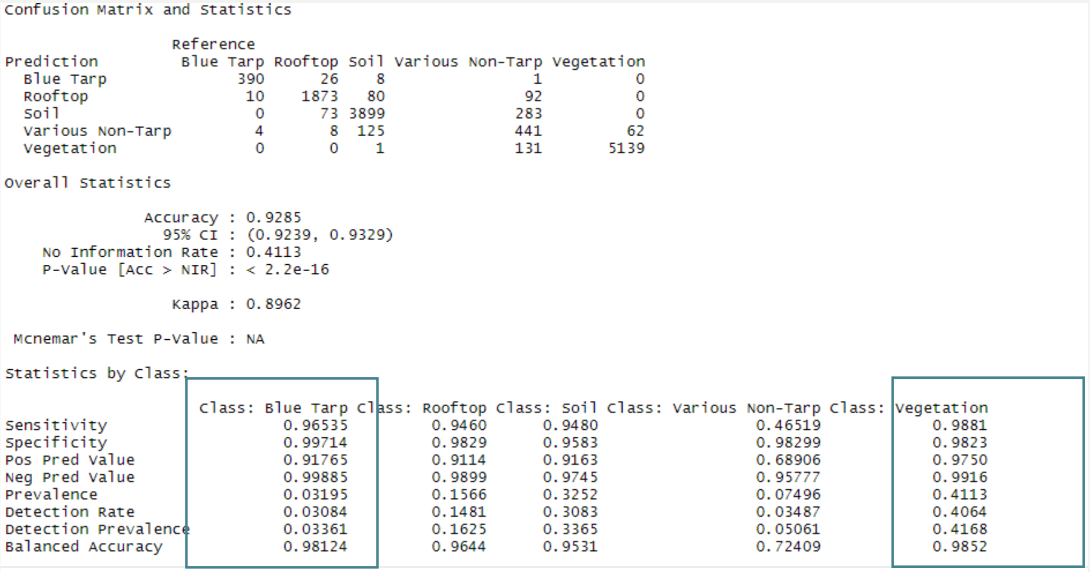


	freq	percentage
Blue Tarp	1618	3.197944
Rooftop	7923	15.65965
Soil	16453	32.519024
Various Non-Tarp	3796	7.502718
Vegetation	20805	41.120664

KNN

- Accuracy: 0.9285 (K=13)
- Sensitivity: 0.9654
- Specificity: 0.9971
- The greatest accuracy occurs at K=13

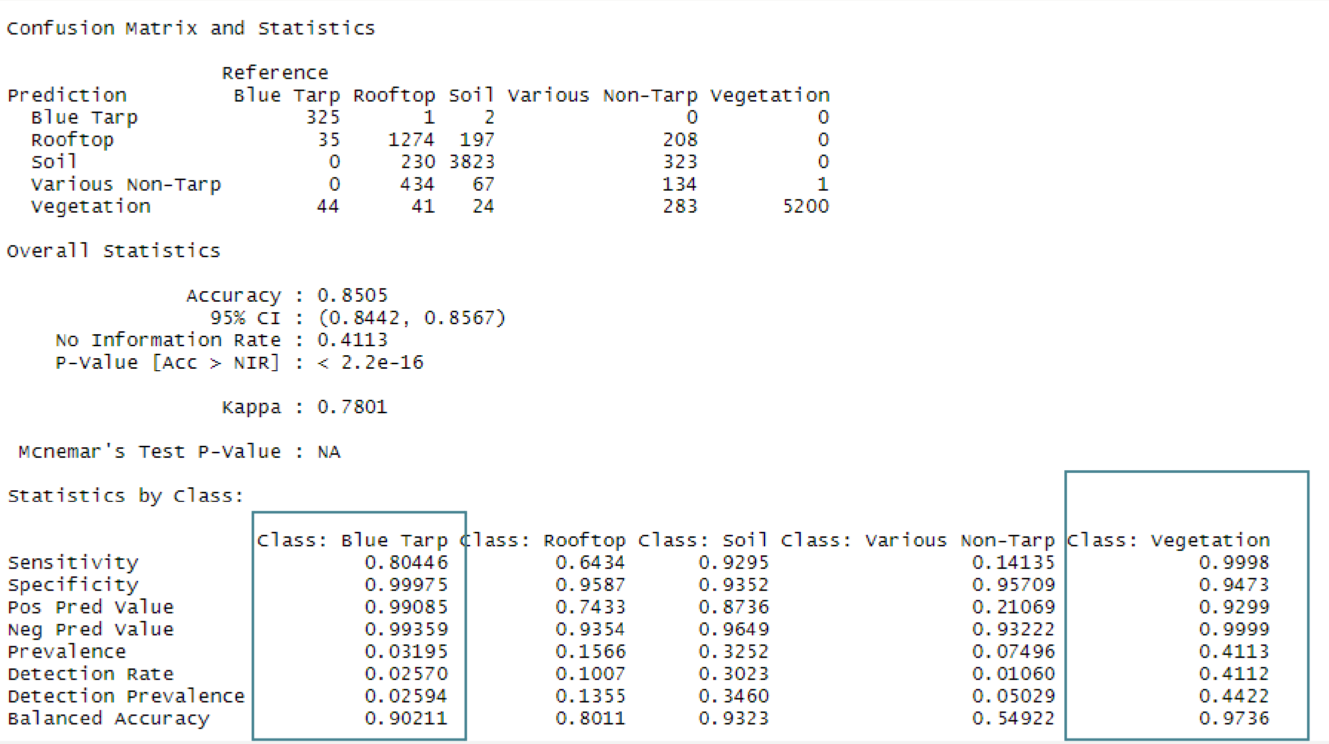




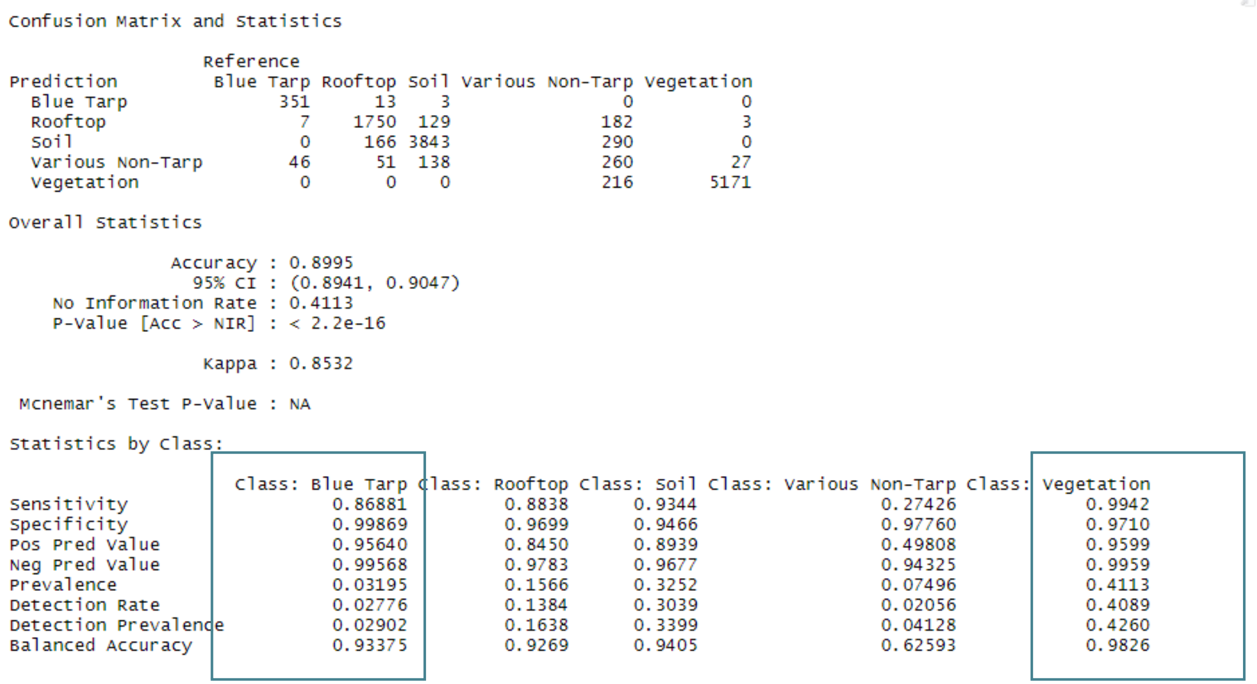
```
Confusion Matrix and Statistics

                  Reference
Prediction          Blue Tarp Rooftop Soil Various Non-Tarp Vegetation
  Blue Tarp               390      26    8                1          0
  Rooftop                  10    1873   80               92          0
  Soil                      0      73 3899              283          0
  Various Non-Tarp          4       8  125              441         62
  Vegetation                0       0    1              131       5139

Overall Statistics

               Accuracy : 0.9285
                 95% CI : (0.9239, 0.9329)
    No Information Rate : 0.4113
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8962

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Blue Tarp Class: Rooftop Class: Soil Class: Various Non-Tarp Class: Vegetation
Sensitivity                   0.96535         0.9460      0.9480                 0.46519            0.9881
Specificity                   0.99714         0.9829      0.9583                 0.98299            0.9823
Pos Pred Value                0.91765         0.9114      0.9163                 0.68906            0.9750
Neg Pred Value                0.99885         0.9899      0.9745                 0.95777            0.9916
Prevalence                    0.03195         0.1566      0.3252                 0.07496            0.4113
Detection Rate                0.03084         0.1481      0.3083                 0.03487            0.4064
Detection Prevalence          0.03361         0.1625      0.3365                 0.05061            0.4168
Balanced Accuracy             0.98124         0.9644      0.9531                 0.72409            0.9852
```

LDA

- Accuracy: 0.8505
- Sensitivity: 0.8044
- Specificity: 0.9997



```
Confusion Matrix and Statistics

                  Reference
Prediction          Blue Tarp Rooftop Soil Various Non-Tarp Vegetation
  Blue Tarp               325       1    2                0          0
  Rooftop                  35    1274  197              208          0
  Soil                      0     230 3823              323          0
  Various Non-Tarp          0     434   67              134          1
  Vegetation               44      41   24              283       5200

Overall Statistics

               Accuracy : 0.8505
                 95% CI : (0.8442, 0.8567)
    No Information Rate : 0.4113
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7801

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Blue Tarp Class: Rooftop Class: Soil Class: Various Non-Tarp Class: Vegetation
Sensitivity                   0.80446         0.6434      0.9295                 0.14135            0.9998
Specificity                   0.99975         0.9587      0.9352                 0.95709            0.9473
Pos Pred Value                0.99085         0.7433      0.8736                 0.21069            0.9299
Neg Pred Value                0.99359         0.9354      0.9649                 0.93222            0.9999
Prevalence                    0.03195         0.1566      0.3252                 0.07496            0.4113
Detection Rate                0.02570         0.1007      0.3023                 0.01060            0.4112
Detection Prevalence          0.02594         0.1355      0.3460                 0.05029            0.4422
Balanced Accuracy             0.90211         0.8011      0.9323                 0.54922            0.9736
```



```
Confusion Matrix and Statistics

                  Reference
Prediction          Blue Tarp Rooftop Soil Various Non-Tarp Vegetation
  Blue Tarp               351      13    3                0          0
  Rooftop                   7    1750  129              182          3
  Soil                      0     166 3843              290          0
  Various Non-Tarp         46      51  138              260         27
  Vegetation                0       0    0              216       5171

Overall Statistics

               Accuracy : 0.8995
                 95% CI : (0.8941, 0.9047)
    No Information Rate : 0.4113
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8532

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Blue Tarp Class: Rooftop Class: Soil Class: Various Non-Tarp Class: Vegetation
Sensitivity                   0.86881         0.8838      0.9344                 0.27426            0.9942
Specificity                   0.99869         0.9699      0.9466                 0.97760            0.9710
Pos Pred Value                0.95640         0.8450      0.8939                 0.49808            0.9599
Neg Pred Value                0.99568         0.9783      0.9677                 0.94325            0.9959
Prevalence                    0.03195         0.1566      0.3252                 0.07496            0.4113
Detection Rate                0.02776         0.1384      0.3039                 0.02056            0.4089
Detection Prevalence          0.02902         0.1638      0.3399                 0.04128            0.4260
Balanced Accuracy             0.93375         0.9269      0.9405                 0.62593            0.9826
```

QDA

Accuracy: 0.8995

Sensitivity: 0.8688

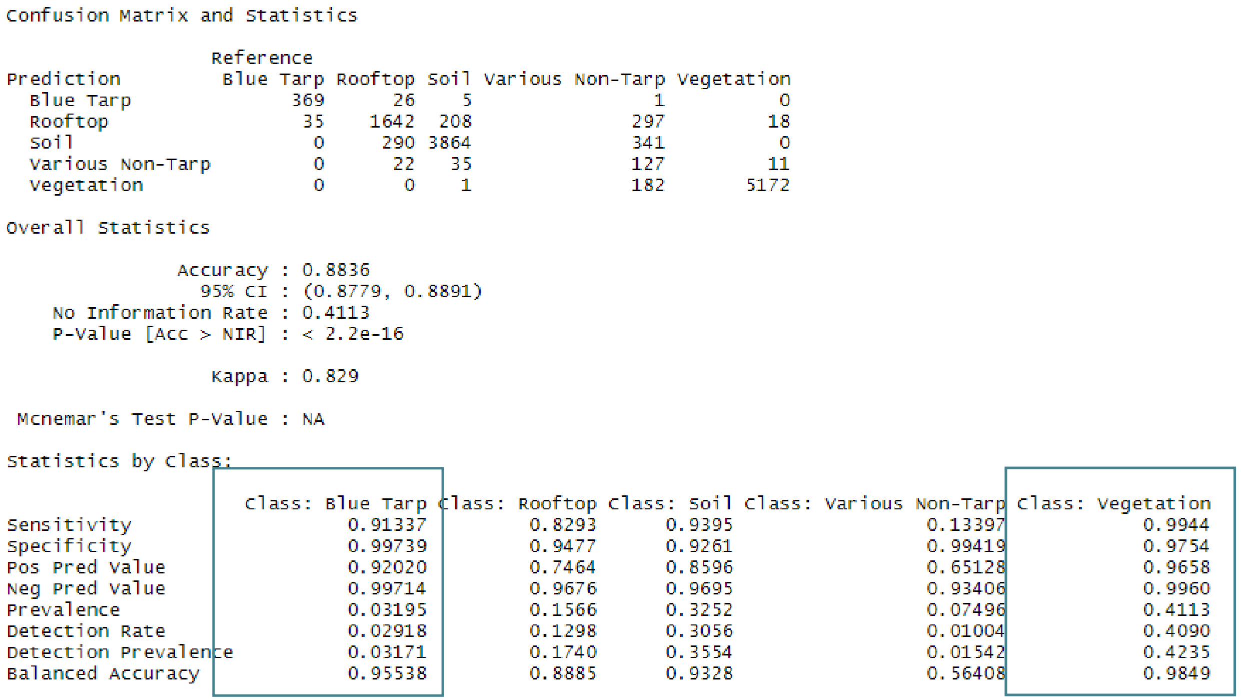
Specificity: 0.9987

LOGISTIC REGRESSION

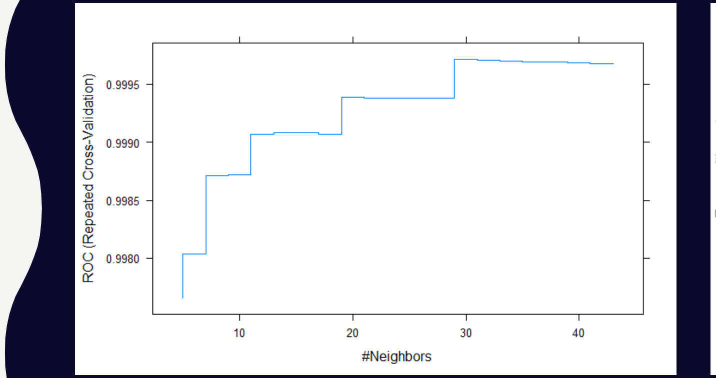
Accuracy: 0.8836

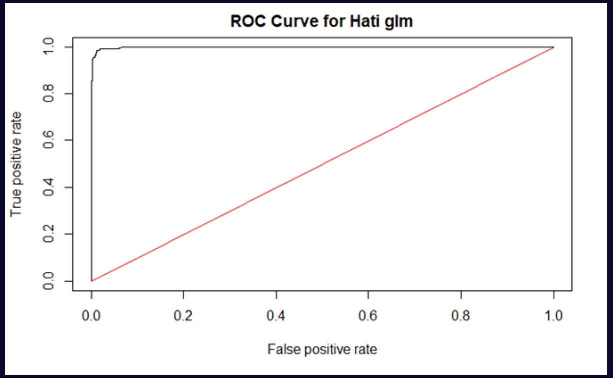
Sensitivity: 0.9134

Specificity: 0.9974



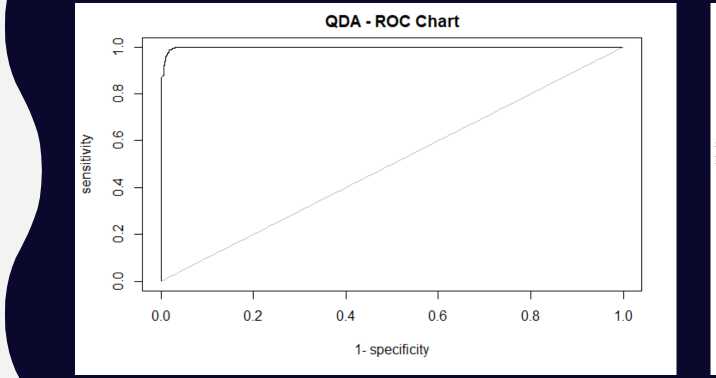
```
Confusion Matrix and Statistics

                  Reference
Prediction         Blue Tarp Rooftop Soil Various Non-Tarp Vegetation
  Blue Tarp              369      26    5                1          0
  Rooftop                 35    1642  208              297         18
  Soil                     0     290 3864              341          0
  Various Non-Tarp         0      22   35              127         11
  Vegetation               0       0    1              182       5172

Overall Statistics

               Accuracy : 0.8836
                 95% CI : (0.8779, 0.8891)
    No Information Rate : 0.4113
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.829

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Blue Tarp Class: Rooftop Class: Soil Class: Various Non-Tarp Class: Vegetation
Sensitivity                   0.91337         0.8293      0.9395                 0.13397            0.9944
Specificity                   0.99739         0.9477      0.9261                 0.99419            0.9754
Pos Pred Value                0.92020         0.7464      0.8596                 0.65128            0.9658
Neg Pred Value                0.99714         0.9676      0.9695                 0.93406            0.9960
Prevalence                    0.03195         0.1566      0.3252                 0.07496            0.4113
Detection Rate                0.02918         0.1298      0.3056                 0.01004            0.4090
Detection Prevalence          0.03171         0.1740      0.3554                 0.01542            0.4235
Balanced Accuracy             0.95538         0.8885      0.9328                 0.56408            0.9849
```

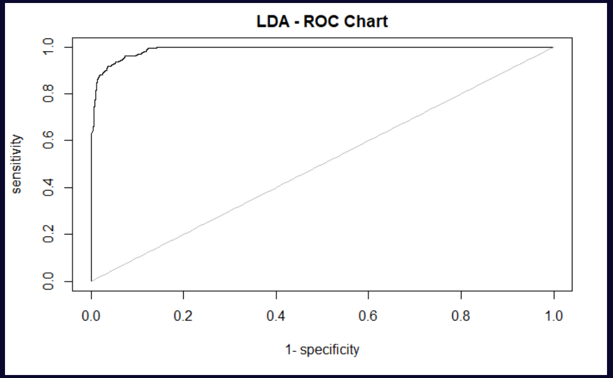




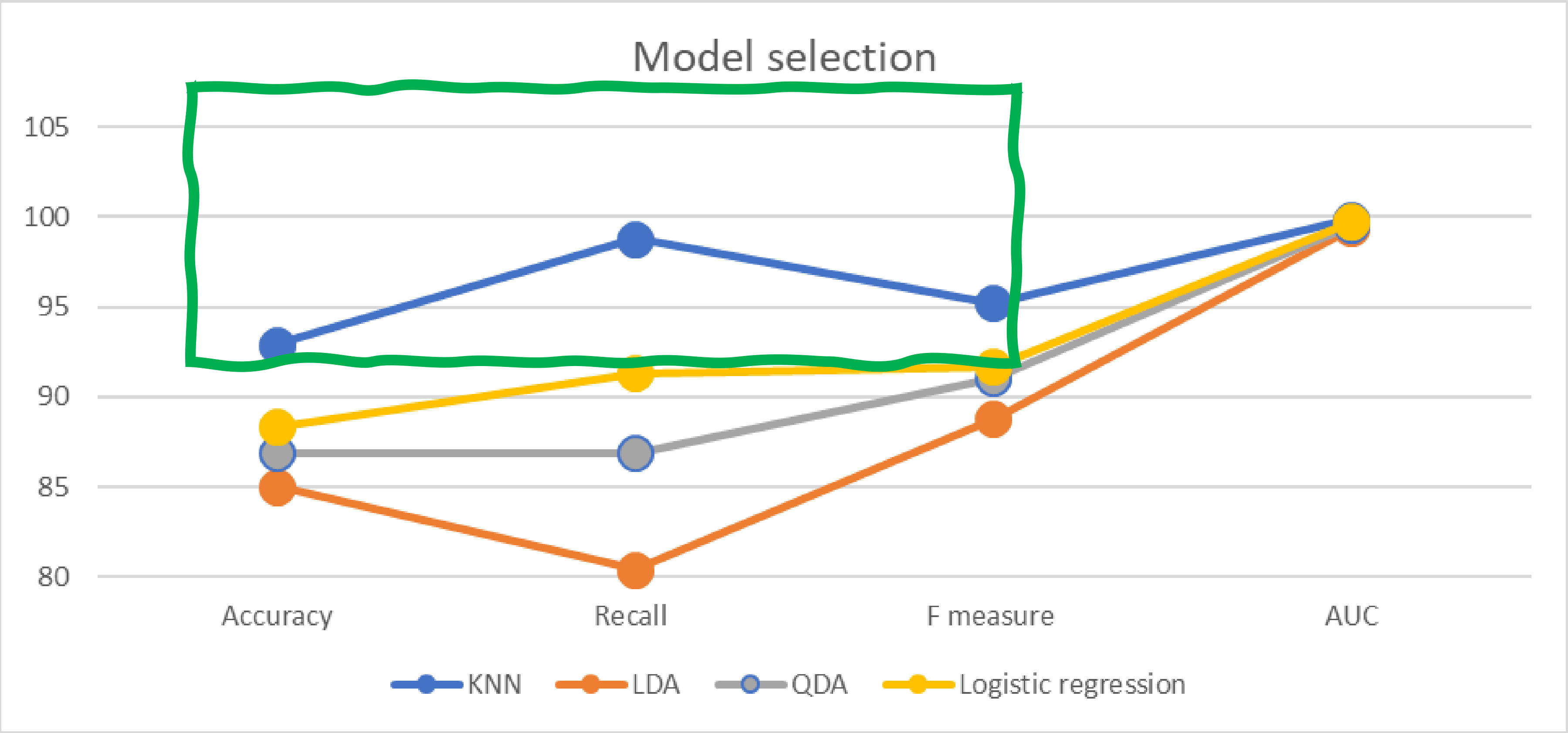
ROC CURVES

- AUC: 0.992 ~ 0.998 obtained from four models





SUMMARY PLOTS



CONCLUSIONS

- Recommend to adopt **KNN** model which has the highest sensitivity rate (96.54%). The reason and the purpose of this study is to predict the "blue tarp" correctly. This means the higher the true positive rate, the greater the accuracy. Since "blue tarp" proportion is only 3%, sensitivity(recall) is adopted as index for model selection.
- Clearly, KNN model shows the best in accuracy, recall, F measure and AUC (K=13)
- The true negative rate (> 99%) are high for all models due to blue tarp being only 3.2% in total.
- Noticeably "Vegetation" is being predicted quite well for all models, KNN, QDA, LDA and logistic regression with sensitivity/specificity ~> 95%-99%. The reason being the proportion of "Vegetation" is 41% among the five classes.