Predicting the occurrence of wildfires with the help remote sensing and weather data using machine learning techniques

GROUP 7

MSC DATA ANALYTICS

BATCH A

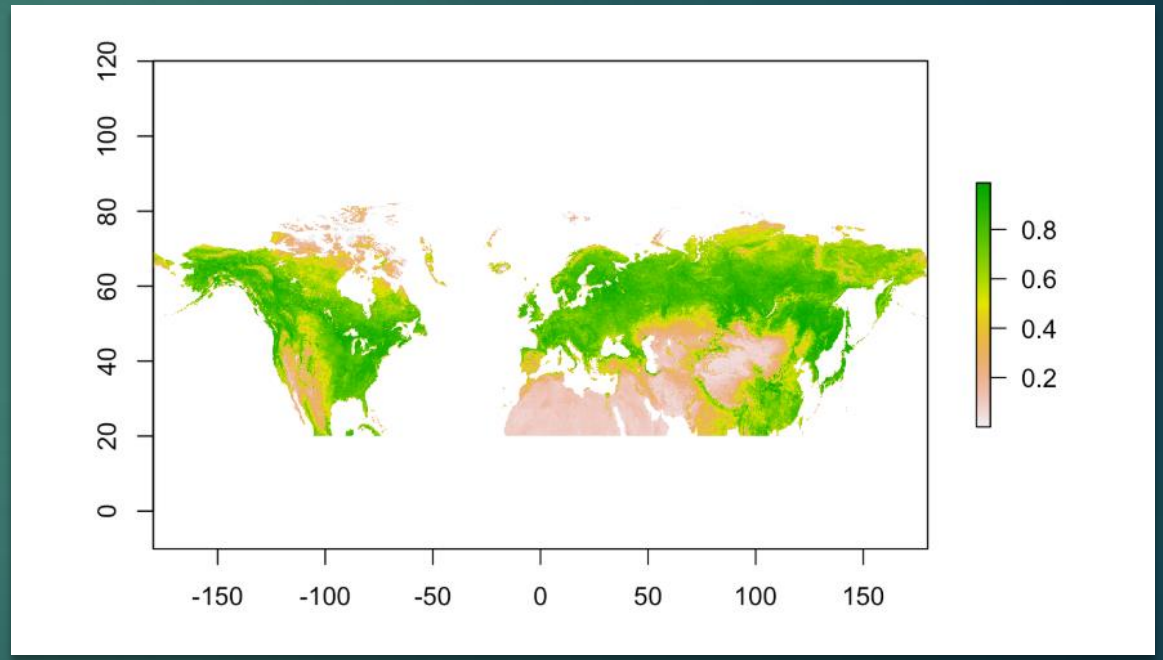
SEPT 2018 - 19

The problem

- Loss of property and life
- Uncertainty in causes of fire – Human, lightning or climatic
- Factors vary region to region

Challenges

- To identify suitable climatic factors for Canada region
- Data processing and merging all the factors with day level granularity and location.
- Extracting NDVI from .nc files



Importance

- Cannot be prevented but can be prepared
- Can prevent loss of life by early warnings
- People can be migrated early

Research Question

Can the prediction of wildfires be improved with the help of weather attributes such as air temperature, humidity and wind in addition to remote-sensing features using machine learning models?

About the Data

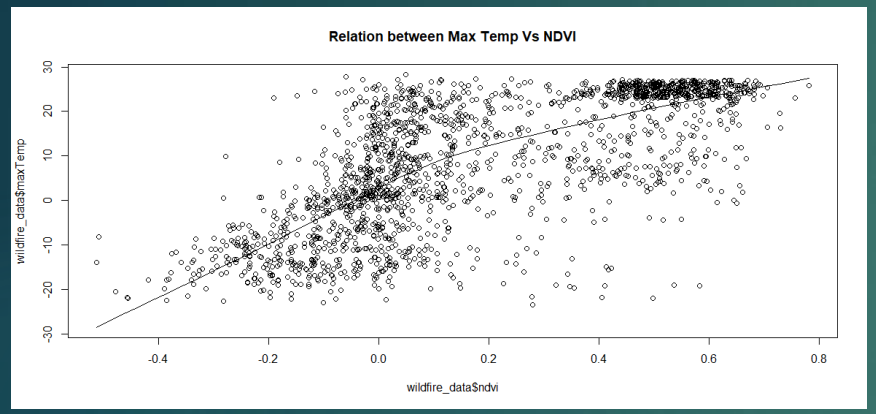
- Wild fire data – from NRC website
- Weather data numeric with lat long from WWO website
- NDVI extracted from nc files using lat long from NCEI
- Huge amount each .nc file was around 50 mb
- Only 2010 data was considered due to such huge amounts

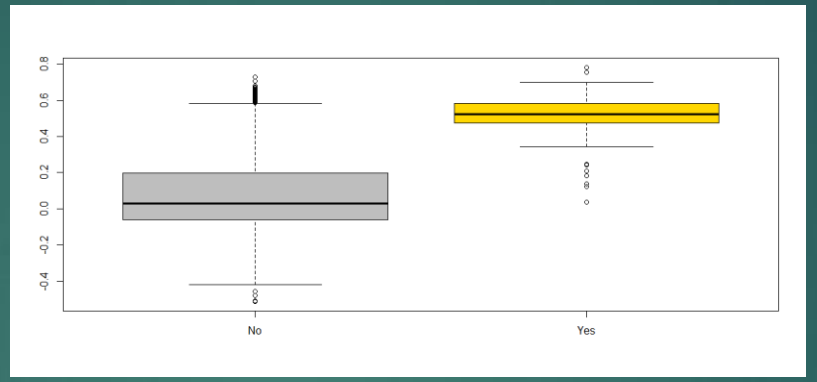
Approach taken

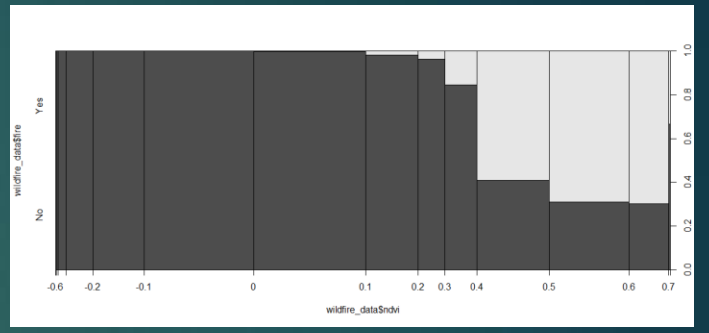
- CRISP-DM Approach
- Exploratory Data Analysis – Class imbalance, univariate, and bivariate analysis.
- Scaling of data
- Correlation matrix
- Random Forest, SVM, ANN, and AutoML

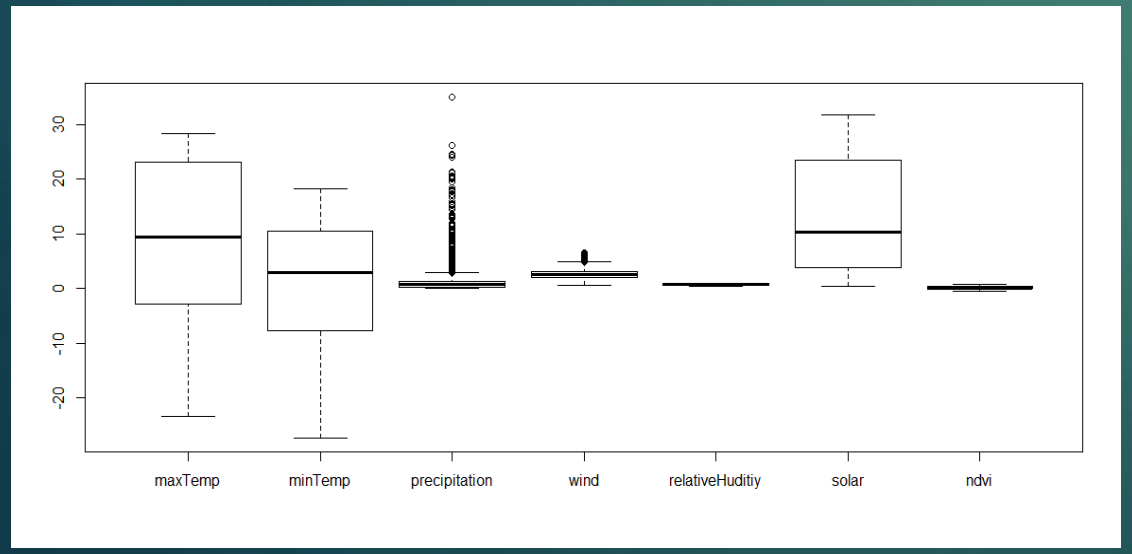
EDA Overview

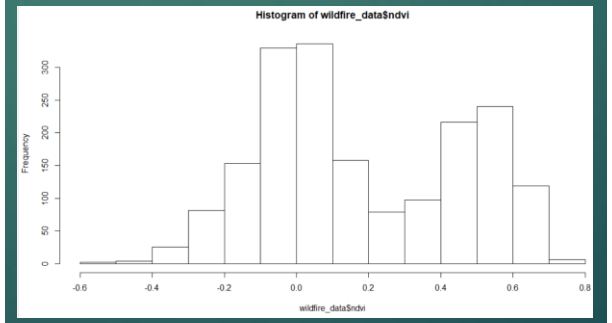
```
> table(wildfire_data$fire)

   No  Yes
1445  401
```

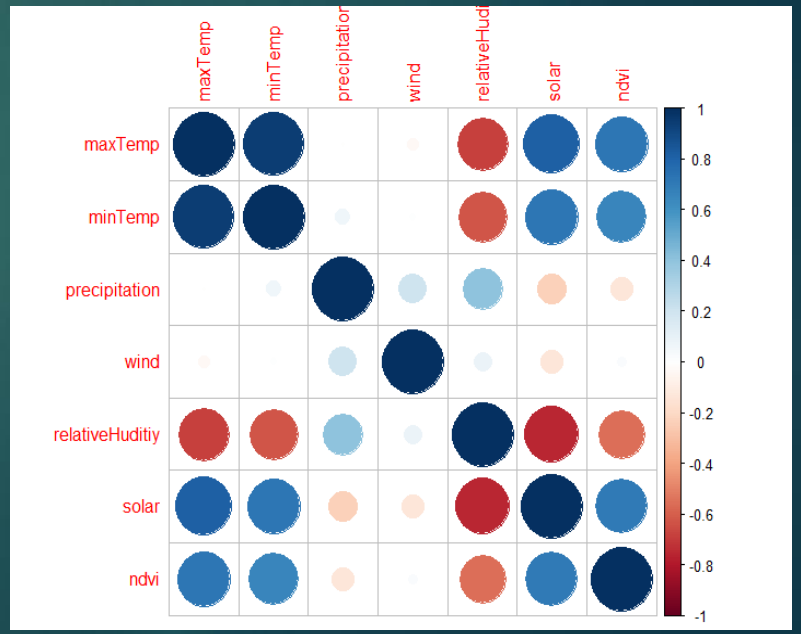








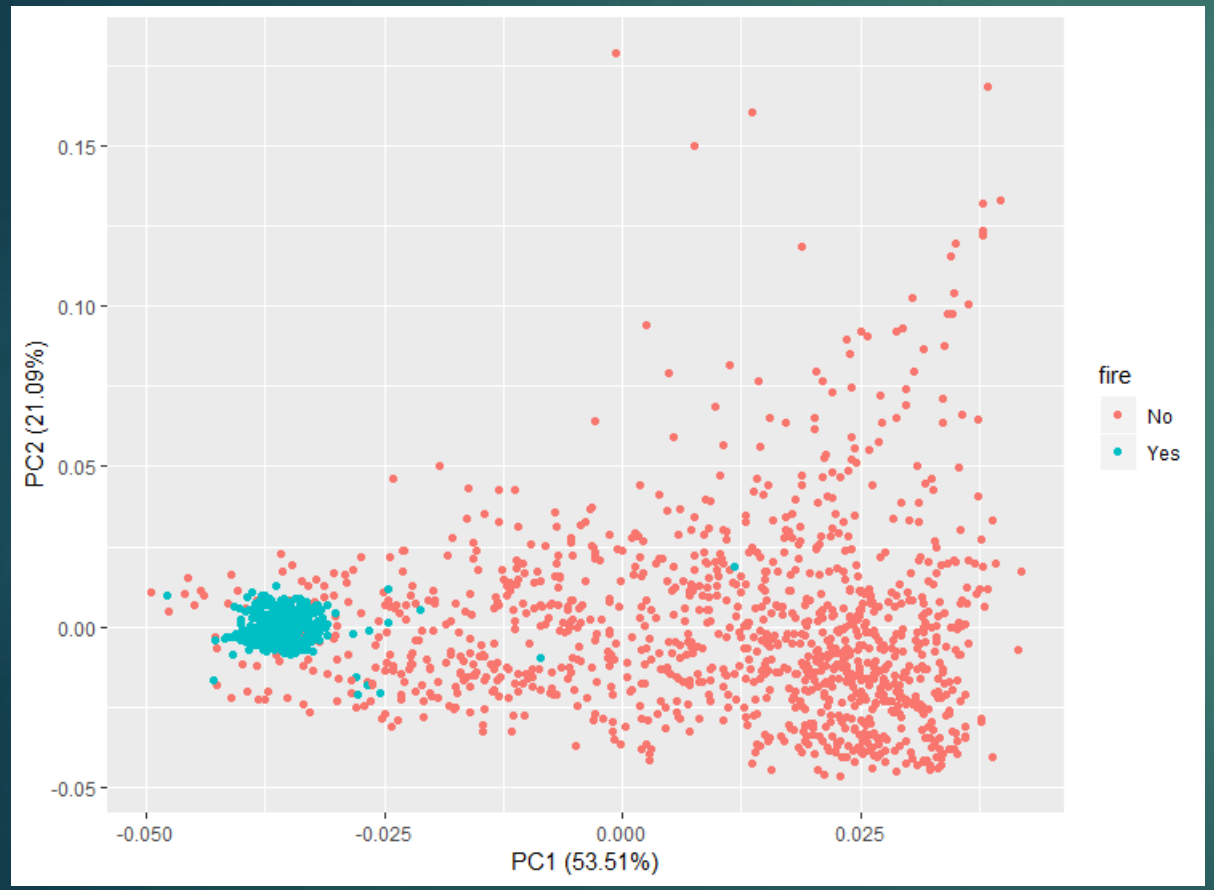


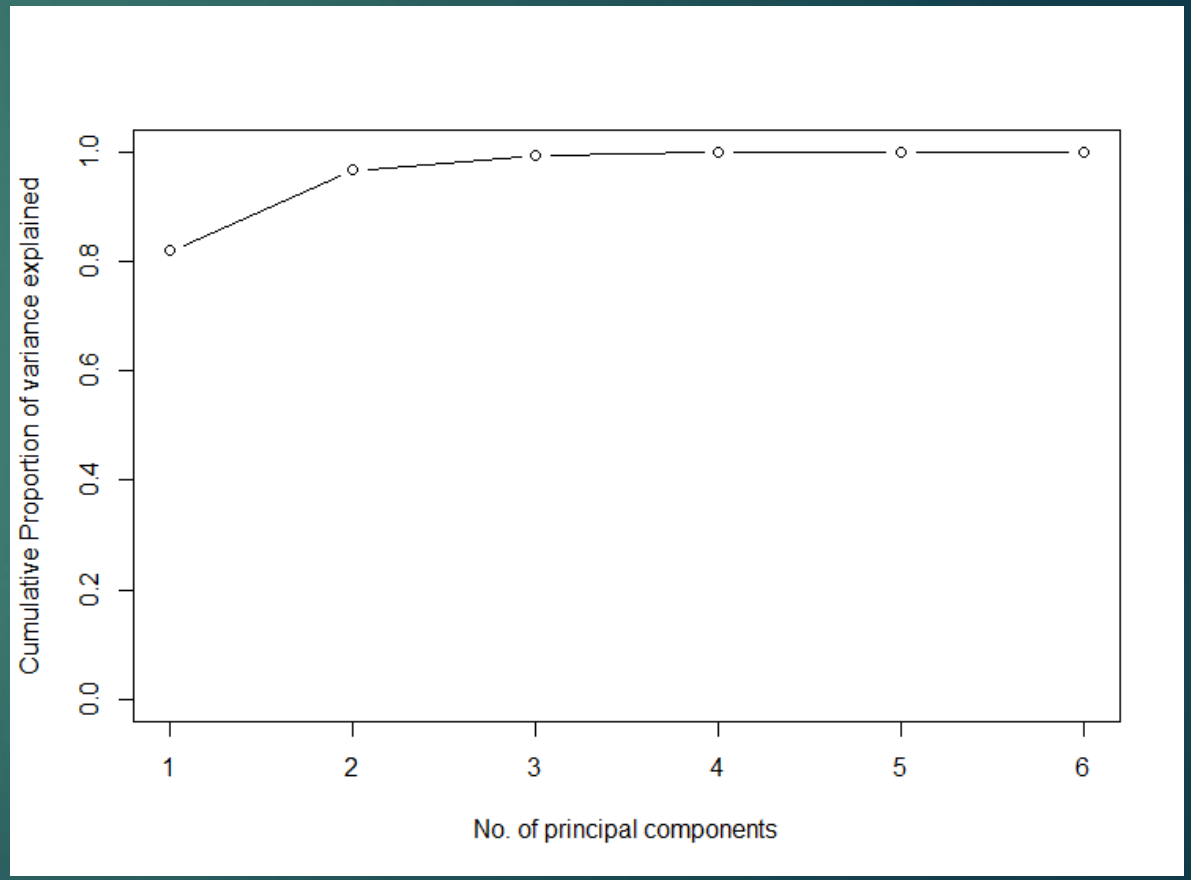


ANN and AutoML

- PCA to check minimum components – 4 from 6 components
- PCA with ANN and activation function – tanh, (160,3), 20 epochs
- PCA with ANN - with hyperparameters settings
- AutoML with PCA
- Training & testing Split - 75/25%

PCA

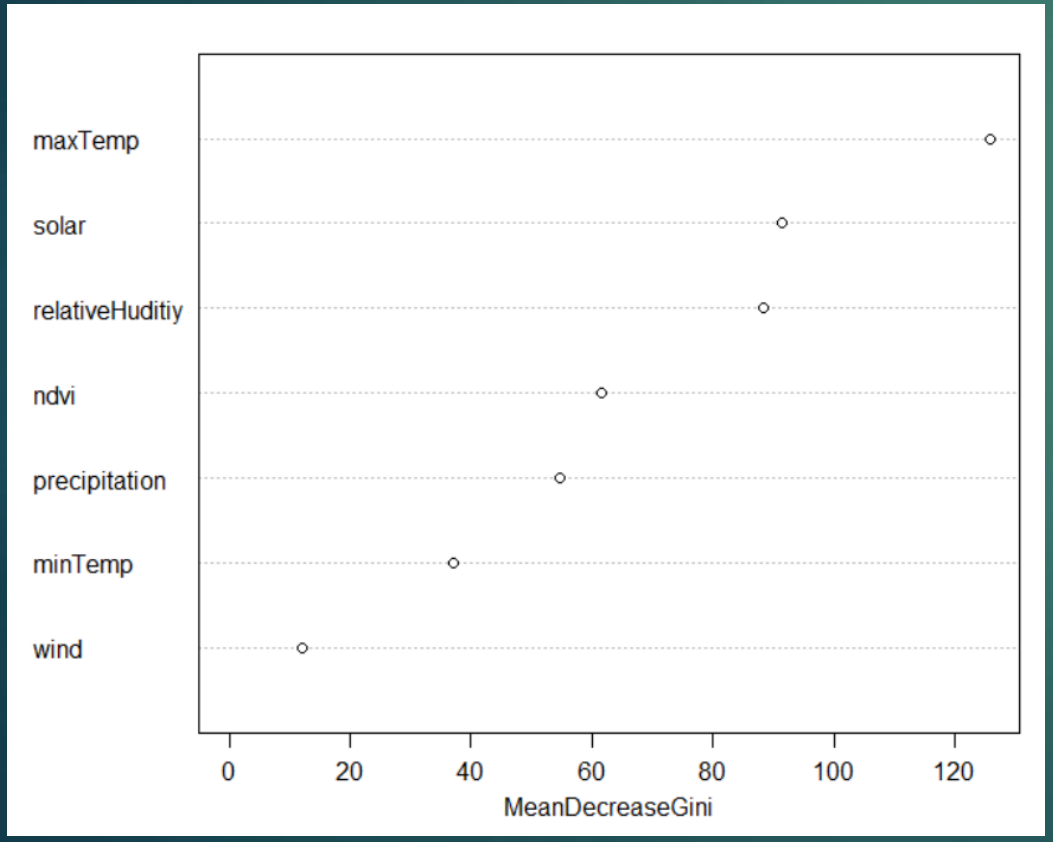

0.15
0.10
0.05
0.00
-0.05
PC2 (21.09%)
-0.050
-0.025
0.000
0.025
PC1 (53.51%)
fire
No
Yes

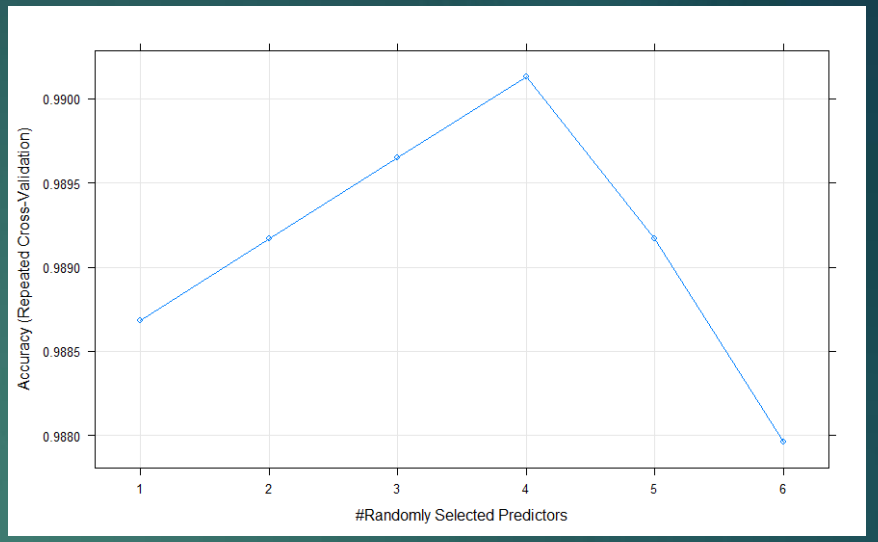

Cumulative Proportion of variance explained
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
No. of principal components

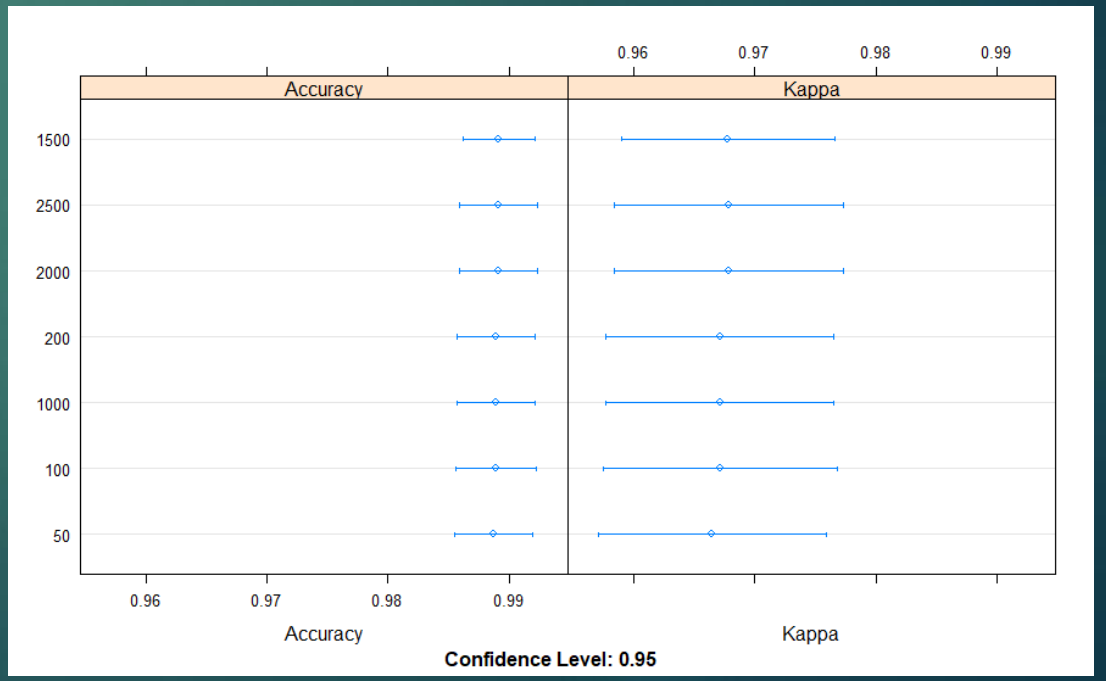
RF and SVM

- Basic RF model with ntree=2000.
- Optmizing parameters of RF by trying different ntree and mtry values.
- Final RF model with ntree=2000 and mtry=4
- SVM with linear, and radial kernel.
- Both SVM models were optimized using n-fold cross validation

RF


maxTemp
solar
relativeHuditiy
ndvi
precipitation
minTemp
wind
0
20
40
60
80
100
120
MeanDecreaseGini


0.9900
0.9895
0.9890
0.9885
0.9880
Accuracy (Repeated Cross-Validation)
1
2
3
4
5
6
#Randomly Selected Predictors


0.96
0.97
0.98
0.99
Accuracy
Kappa
1500
2500
2000
200
1000
100
50
0.96
0.97
0.98
0.99
Accuracy
Kappa
Confidence Level: 0.95

Results

COMPARISON OF MODELS

Methodology	Accuracy	Sensitivity	Specificity
RF-basic	0.9957	0.98	1
RF-optimized	0.9935	0.98	0.9972
SVM-linear	0.9761	0.9700	0.9778
SVM-radial	0.9848	0.98	0.9861
ANN	0.9805	0.9945	0.963
AutoML	0.9706	0.9832	0.9265
Base Model	0.9832	0.98	0.9797

Future work

- To predict the spread area
- Human factor not taken into consideration
- Proximity to civilization as a factor should be taken into consideration