CDA 541 STATISTICAL DATA MINING I

Final Project

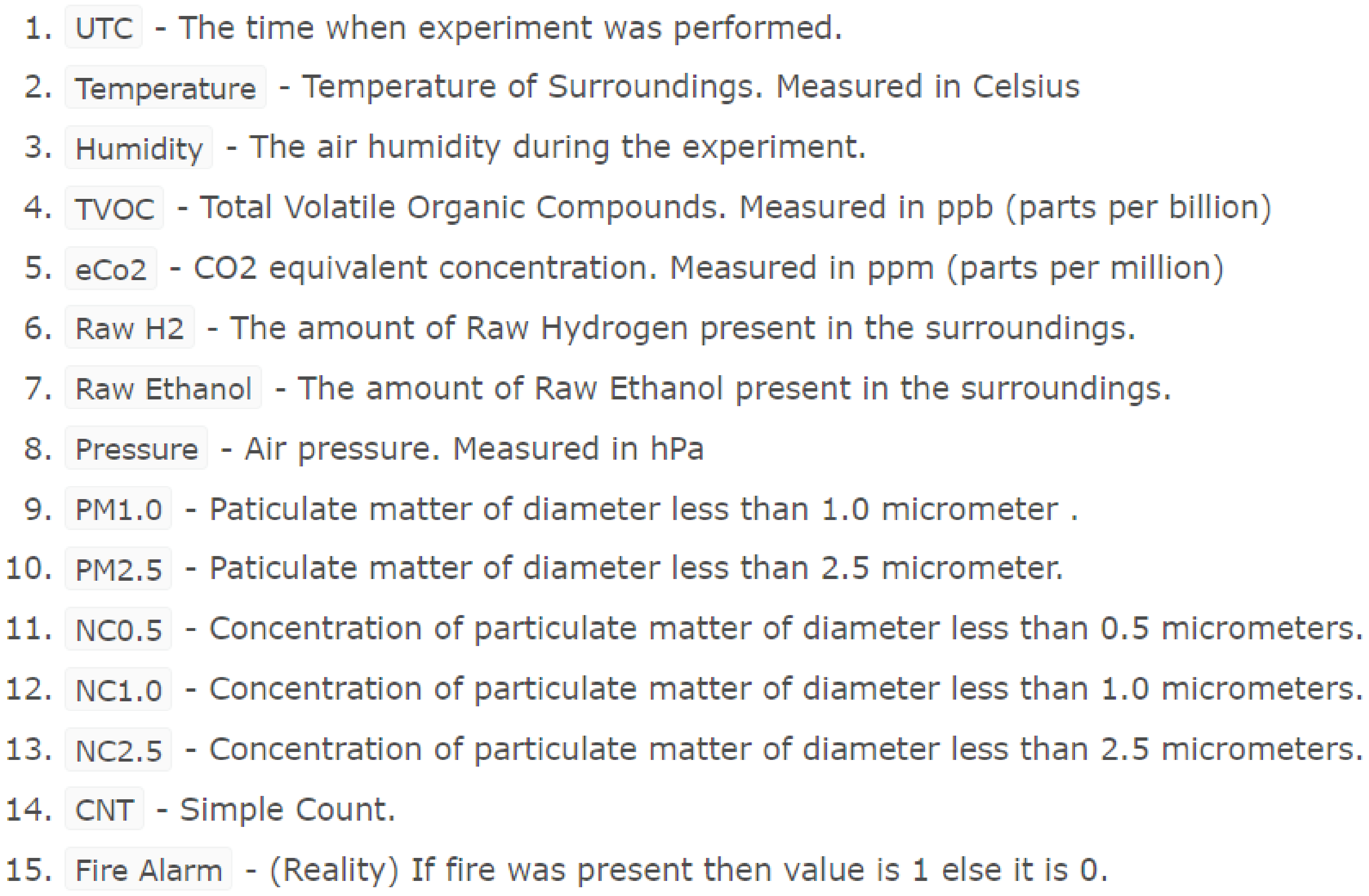
University at Buffalo
School of Engineering and Applied Sciences

Predicting Fire alarm Trigger using Smoke Detection Data

The goal of this project is to predict the trigger of the fire alarm using the smoke data. We are having 60k observations and 15 features in the data to predict the outcome. These features include the amount of chemical compounds present in the air along with some properties like Humidity , Pressure etc.

We are going to build multiple classification models, check their performance on this dataset , to determine the best model for this data. Classification Models include - Logistic Regression , KNN , Decision Tree , Random Forest , ADA Boost , Gradient Boost . Best model will be proposed based on the train and test error and accuracy of the model.

Dataset Overview



1. `UTC` - The time when experiment was performed.
2. `Temperature` - Temperature of Surroundings. Measured in Celsius
3. `Humidity` - The air humidity during the experiment.
4. `TVOC` - Total Volatile Organic Compounds. Measured in ppb (parts per billion)
5. `eCo2` - CO2 equivalent concentration. Measured in ppm (parts per million)
6. `Raw H2` - The amount of Raw Hydrogen present in the surroundings.
7. `Raw Ethanol` - The amount of Raw Ethanol present in the surroundings.
8. `Pressure` - Air pressure. Measured in hPa
9. `PM1.0` - Paticulate matter of diameter less than 1.0 micrometer .
10. `PM2.5` - Paticulate matter of diameter less than 2.5 micrometer.
11. `NC0.5` - Concentration of particulate matter of diameter less than 0.5 micrometers.
12. `NC1.0` - Concentration of particulate matter of diameter less than 1.0 micrometers.
13. `NC2.5` - Concentration of particulate matter of diameter less than 2.5 micrometers.
14. `CNT` - Simple Count.
15. `Fire Alarm` - (Reality) If fire was present then value is 1 else it is 0.

Data Exploration

There are 62,630 observations and 15 predictor variables and 1 response variables . All variables are either Integer or Float and are continuous variables . There are no categorical variables in this dataset . There are no rows which have null values and no duplicate records as well . This dataset is very clean we can proceed with EDA .

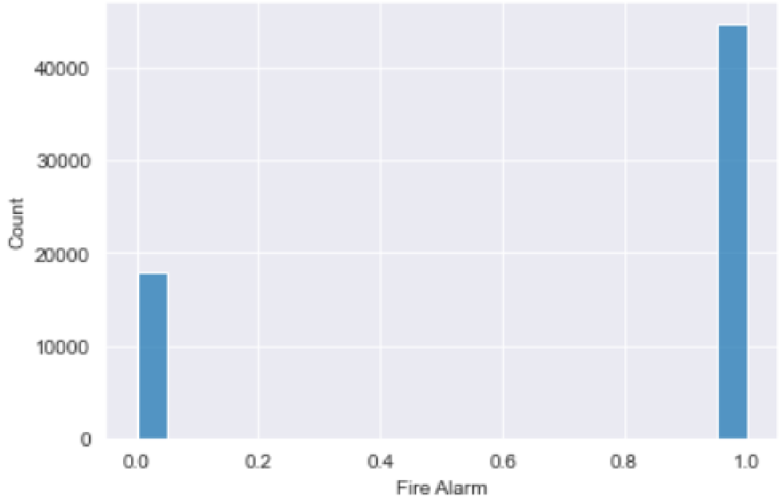
There are some columns which we can drop in the data .

- UTC - This column tells us the time which data point is recorded , which has no effect on the response variable
- Unnamed: 0 , CNT - These are index columns , so we can remove these as well

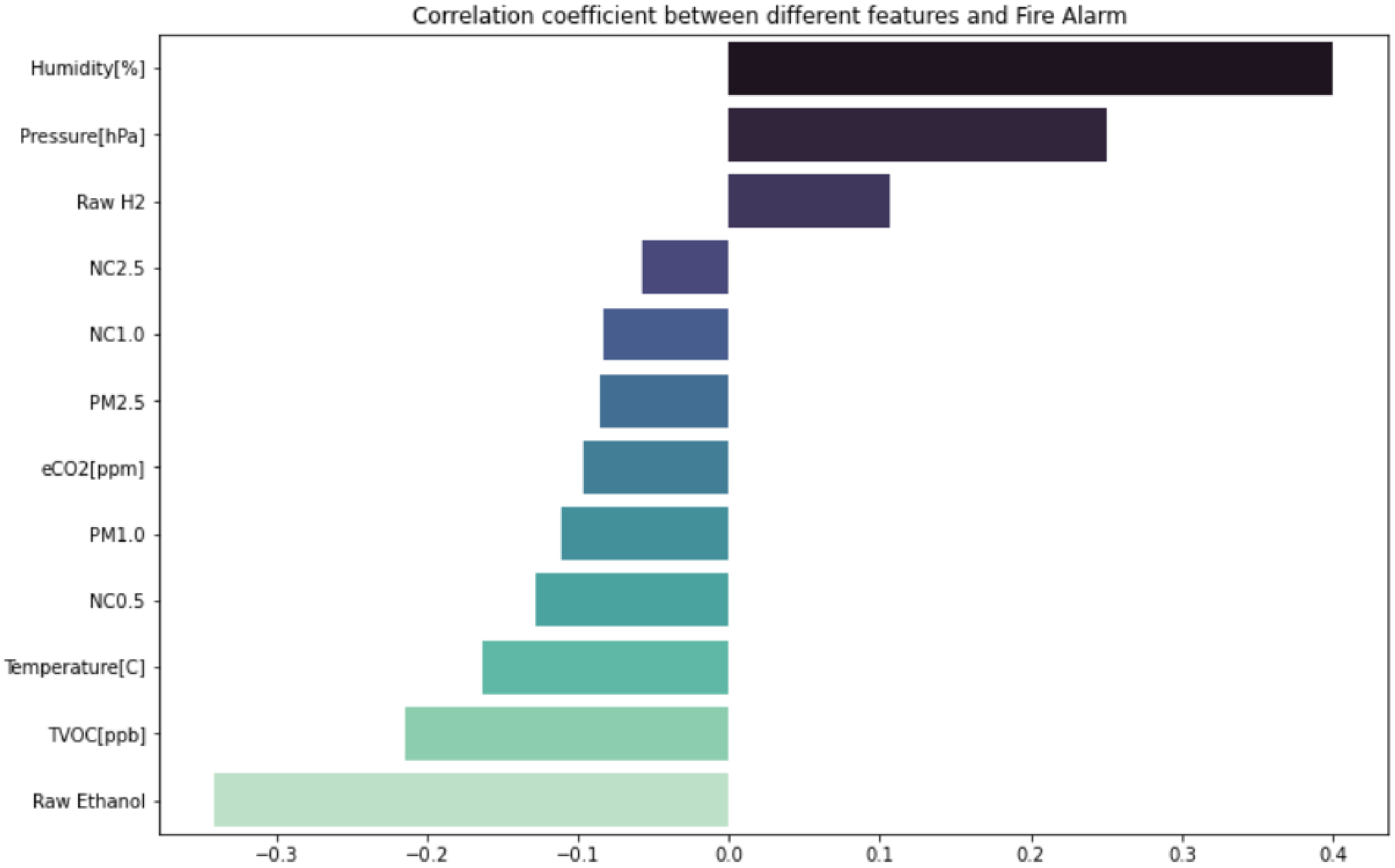
Temperature, PM , NC values are on different scales . We need to scale the data when we are modelling

EDA

- From 60k observations we see only 18k observations belong ‘No Fire’ and remaining 40k observations ‘Yes Fire’
- This dataset looks imbalanced, but we will continue with our analysis and see if this is having any effect on the models. If we see underfit and most of ‘No Fire’ data is predicted as ‘Yes’ we will balance the data and build models again.

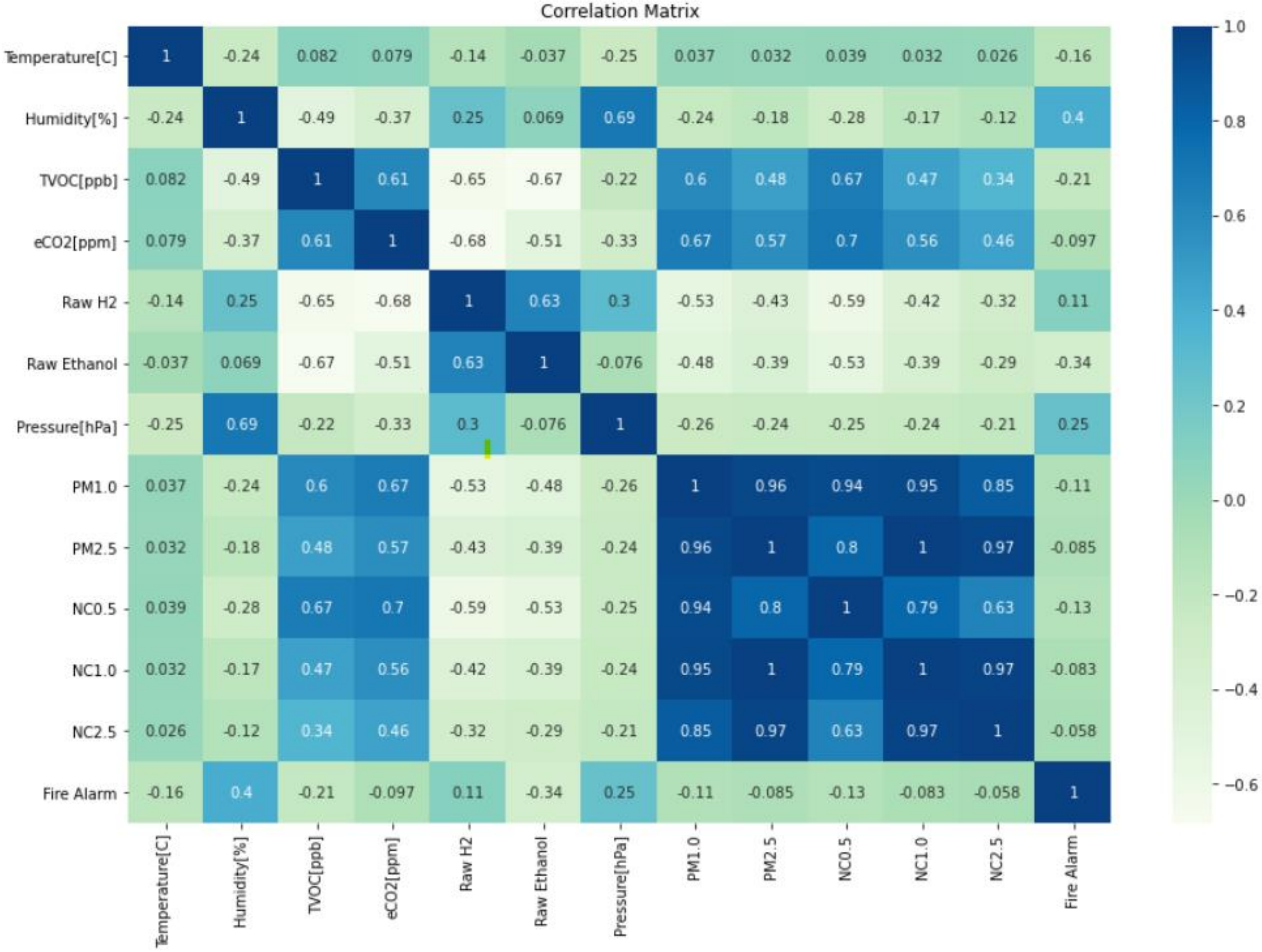


We will check the correlation of response variable with the predictor variables



Humidity and Pressure are having high positive correlation with the response . Raw Ethanol and TVOC are having high negative correlation.

Correlation matrix to see how predictor variables interact with each other



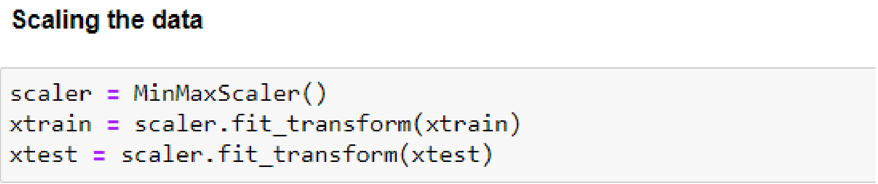
	Temperature[C]	Humidity[%]	TVOC[ppb]	eCO2[ppm]	Raw H2	Raw Ethanol	Pressure[hPa]	PM1.0	PM2.5	NC0.5	NC1.0	NC2.5	Fire Alarm
Temperature[C]	1	-0.24	0.082	0.079	-0.14	-0.037	-0.25	0.037	0.032	0.039	0.032	0.026	-0.16
Humidity[%]	-0.24	1	-0.49	-0.37	0.25	0.069	0.69	-0.24	-0.18	-0.28	-0.17	-0.12	0.4
TVOC[ppb]	0.082	-0.49	1	0.61	-0.65	-0.67	-0.22	0.6	0.48	0.67	0.47	0.34	-0.21
eCO2[ppm]	0.079	-0.37	0.61	1	-0.68	-0.51	-0.33	0.67	0.57	0.7	0.56	0.46	-0.097
Raw H2	-0.14	0.25	-0.65	-0.68	1	0.63	0.3	-0.53	-0.43	-0.59	-0.42	-0.32	0.11
Raw Ethanol	-0.037	0.069	-0.67	-0.51	0.63	1	-0.076	-0.48	-0.39	-0.53	-0.39	-0.29	-0.34
Pressure[hPa]	-0.25	0.69	-0.22	-0.33	0.3	-0.076	1	-0.26	-0.24	-0.25	-0.24	-0.21	0.25
PM1.0	0.037	-0.24	0.6	0.67	-0.53	-0.48	-0.26	1	0.96	0.94	0.95	0.85	-0.11
PM2.5	0.032	-0.18	0.48	0.57	-0.43	-0.39	-0.24	0.96	1	0.8	1	0.97	-0.085
NC0.5	0.039	-0.28	0.67	0.7	-0.59	-0.53	-0.25	0.94	0.8	1	0.79	0.63	-0.13
NC1.0	0.032	-0.17	0.47	0.56	-0.42	-0.39	-0.24	0.95	1	0.79	1	0.97	-0.083
NC2.5	0.026	-0.12	0.34	0.46	-0.32	-0.29	-0.21	0.85	0.97	0.63	0.97	1	-0.058
Fire Alarm	-0.16	0.4	-0.21	-0.097	0.11	-0.34	0.25	-0.11	-0.085	-0.13	-0.083	-0.058	1

We could see all PM and NC values are very highly correlated Humidity and Pressure are moderately correlated as well.

Modelling

Scaling the data

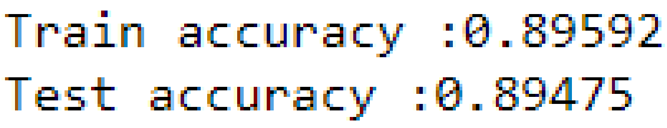
We have used MinMaxScaler() from sklearn to transform the features to be in range 0 to 1



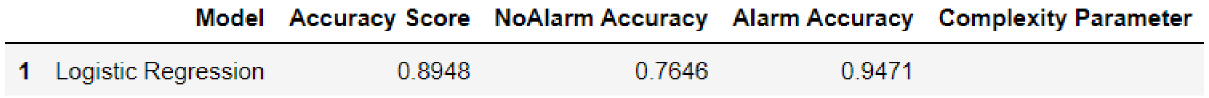
Scaling the data

```
scaler = MinMaxScaler()
xtrain = scaler.fit_transform(xtrain)
xtest = scaler.fit_transform(xtest)
```

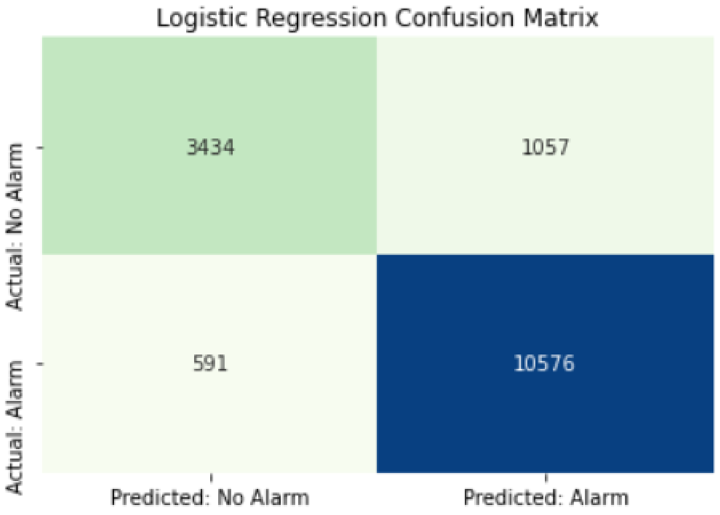
Logistic Regression Results



```
Train accuracy :0.89592
Test accuracy :0.89475
```



	Model	Accuracy Score	NoAlarm Accuracy	Alarm Accuracy	Complexity Parameter
1	Logistic Regression	0.8948	0.7646	0.9471	

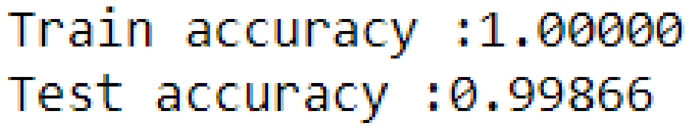


Logistic Regression Confusion Matrix

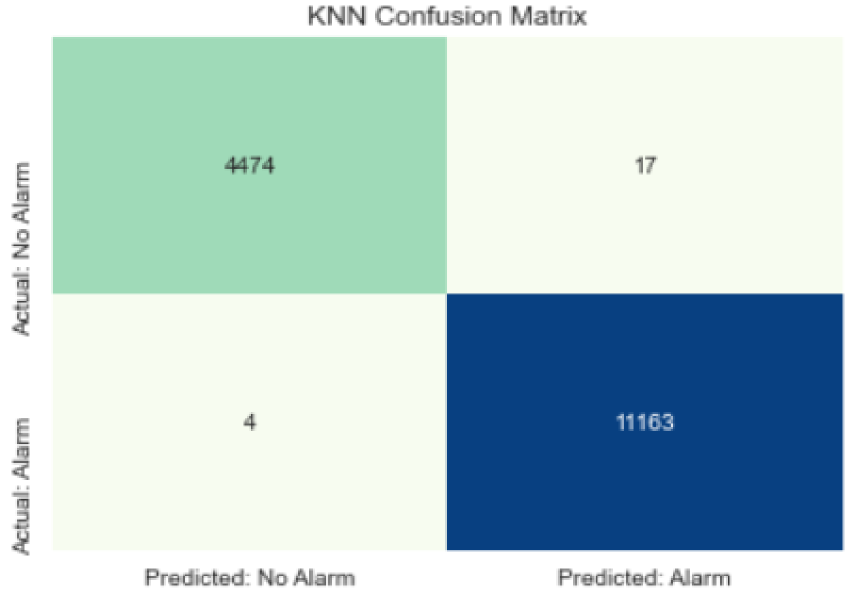
	Predicted: No Alarm	Predicted: Alarm
Actual: No Alarm	3434	1057
Actual: Alarm	591	10576

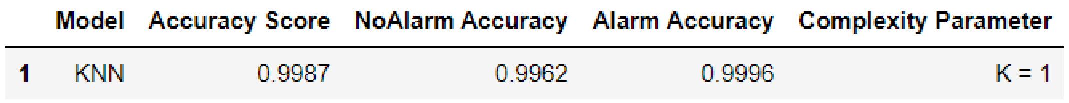
KNN

For KNN , we could see the best accuracy for training and test data for K=1



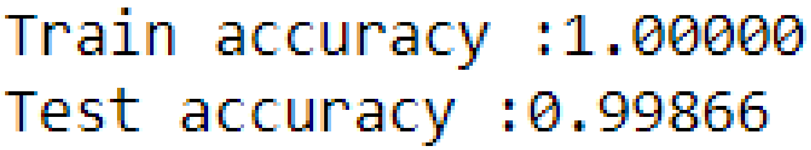
```
Train accuracy :1.00000
Test accuracy :0.99866
```



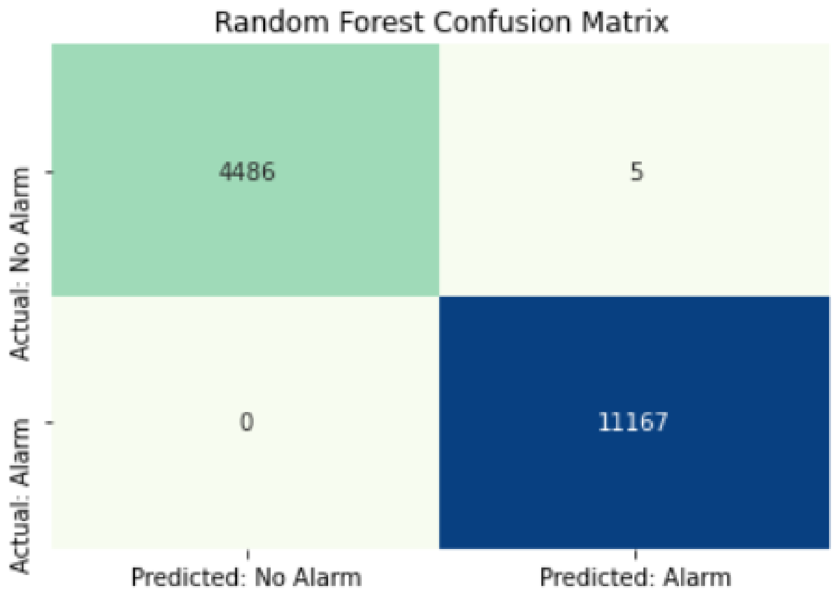


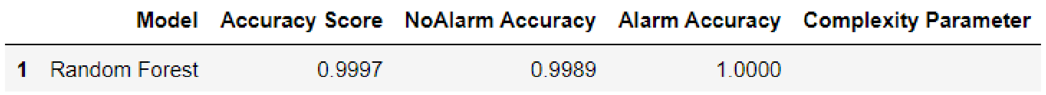
	Model	Accuracy Score	NoAlarm Accuracy	Alarm Accuracy	Complexity Parameter
1	KNN	0.9987	0.9962	0.9996	K = 1

Random Forest



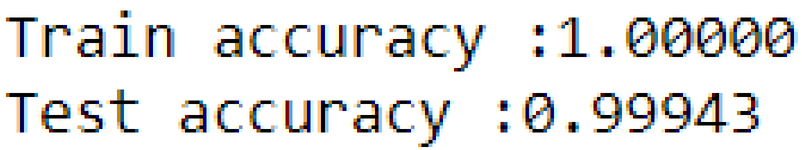
```
Train accuracy :1.00000
Test accuracy :0.99866
```



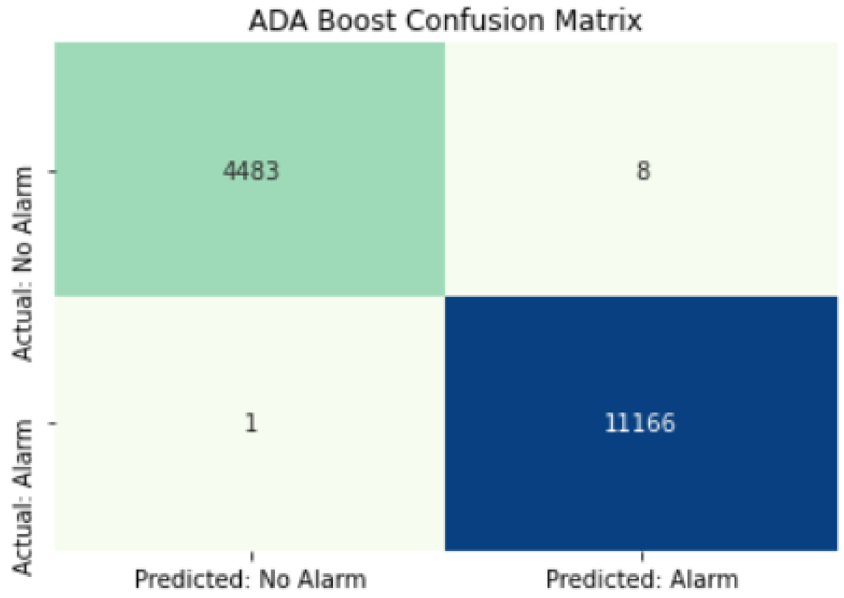


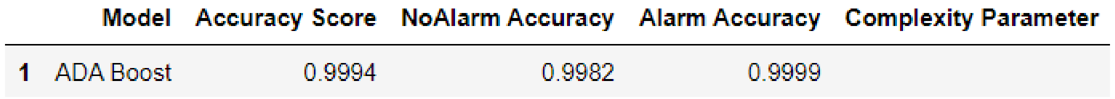
	Model	Accuracy Score	NoAlarm Accuracy	Alarm Accuracy	Complexity Parameter
1	Random Forest	0.9997	0.9989	1.0000	

ADA Boost



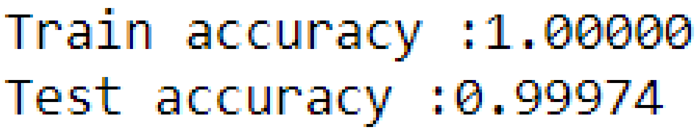
```
Train accuracy :1.00000
Test accuracy :0.99943
```



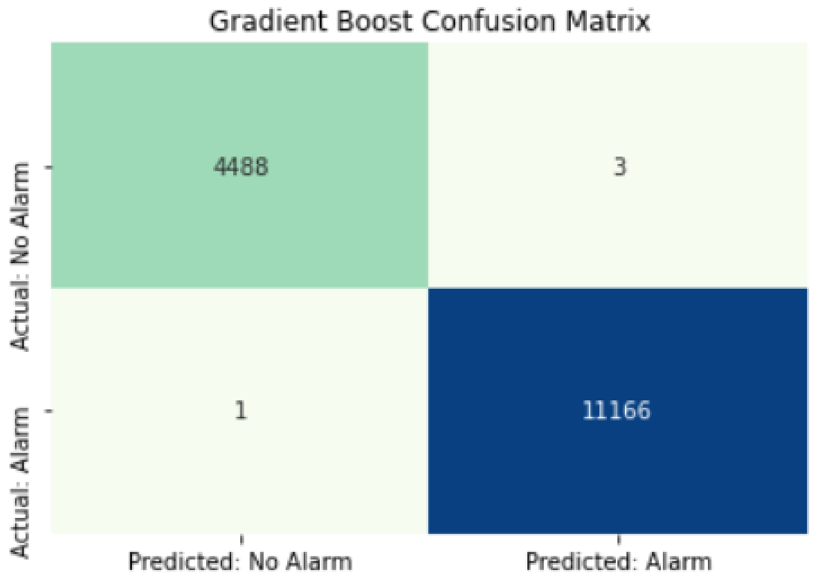


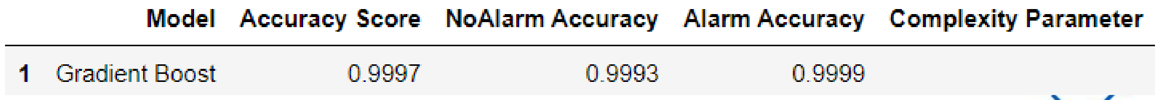
	Model	Accuracy Score	NoAlarm Accuracy	Alarm Accuracy	Complexity Parameter
1	ADA Boost	0.9994	0.9982	0.9999	

Gradient Boost



```
Train accuracy :1.00000
Test accuracy :0.99974
```



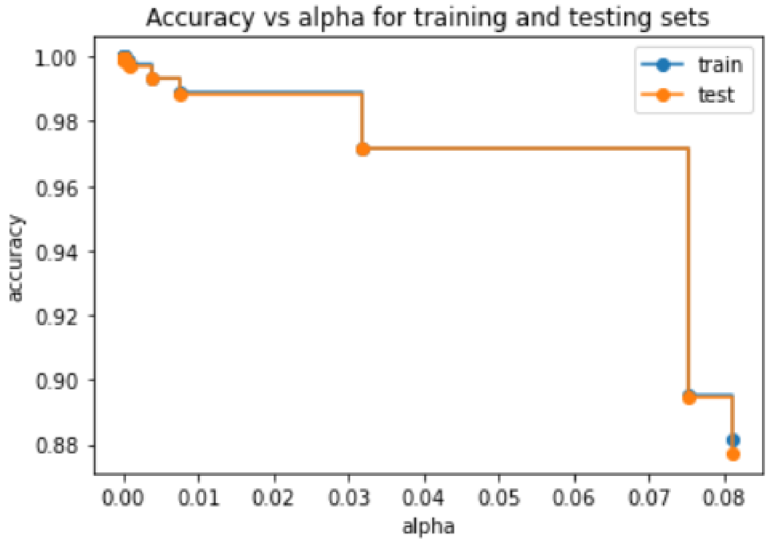


	Model	Accuracy Score	NoAlarm Accuracy	Alarm Accuracy	Complexity Parameter
1	Gradient Boost	0.9997	0.9993	0.9999	

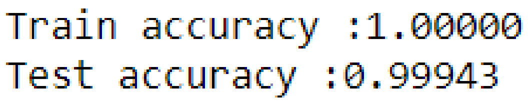
Decision Tree

Finding out the Cost complexity parameter to prune the tree

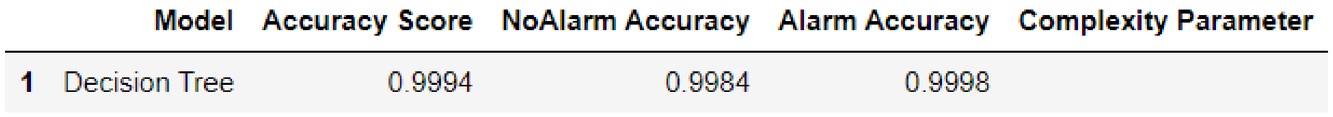
From the graph we could see at cp = 0 we could see best accuracy for both train and test sets . So we are setting cp to 0 , when building our decision tree



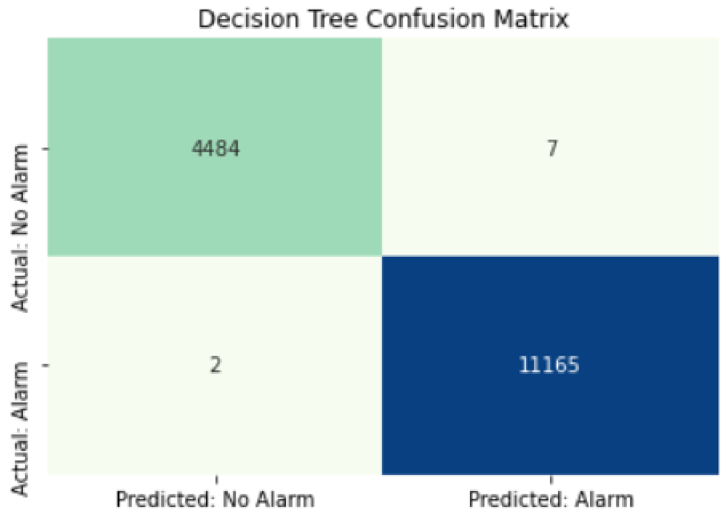
Results



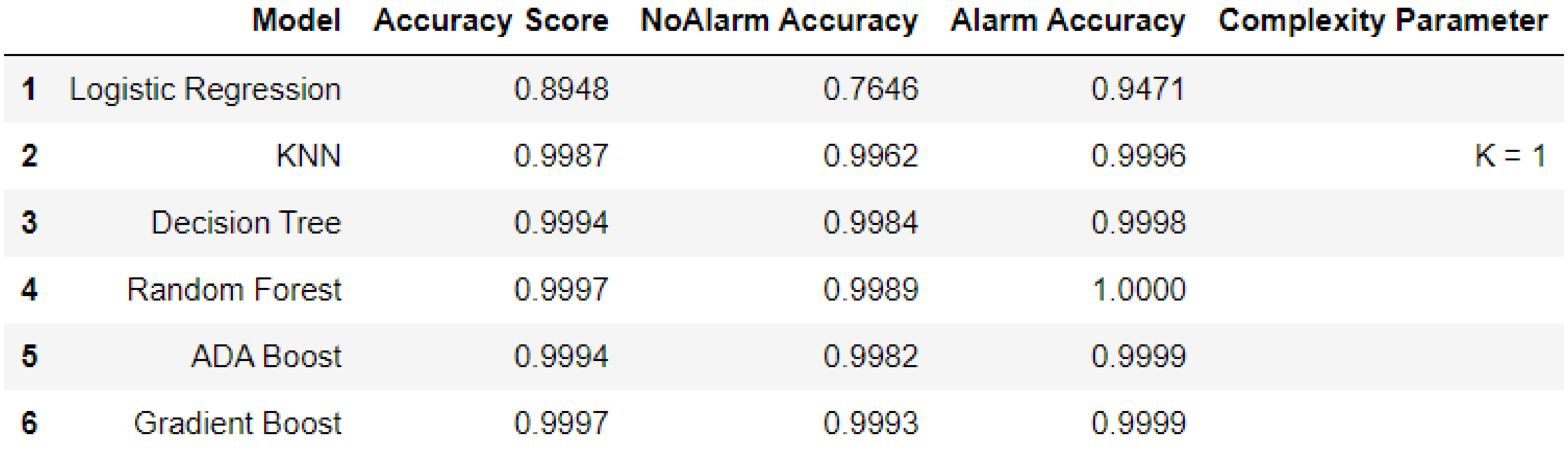
```
Train accuracy :1.00000
Test accuracy :0.99943
```



	Model	Accuracy Score	NoAlarm Accuracy	Alarm Accuracy	Complexity Parameter
1	Decision Tree	0.9994	0.9984	0.9998	



Results Summary



	Model	Accuracy Score	NoAlarm Accuracy	Alarm Accuracy	Complexity Parameter
1	Logistic Regression	0.8948	0.7646	0.9471	
2	KNN	0.9987	0.9962	0.9996	K = 1
3	Decision Tree	0.9994	0.9984	0.9998	
4	Random Forest	0.9997	0.9989	1.0000	
5	ADA Boost	0.9994	0.9982	0.9999	
6	Gradient Boost	0.9997	0.9993	0.9999	

We can see the best accuracy for both ‘Random Forest’ and ‘Gradient Boosting’ models. But if we consider the time complexity of the model as well , ‘Random Forest’ will be the best model for this dataset.

Logistic regression performed poorly when compared. We did not balance the data , but we are able to see good results.

THANK YOU

Name	Email Id
Soujan Vakulabharanam	soujanva@buffalo.edu
Bharanidhar Reddy Marthala	bmarthal@buffalo.edu
Sunil Kumar Madavaram	smadavar@buffalo.edu

University at Buffalo
School of Engineering and Applied Sciences