# **Microsoft : Classifying Cybersecurity Incidents with Machine Learning**

# Missing values count column wise

|  |  |  |  |
| --- | --- | --- | --- |
|  | **NaN Count** | **NaN Percentage** | **Duplicates** |
| **Unnamed: 0** | 0 | 0 | FALSE |
| **Id** | 0 | 0 | TRUE |
| **OrgId** | 0 | 0 | TRUE |
| **IncidentId** | 0 | 0 | TRUE |
| **AlertId** | 0 | 0 | TRUE |
| **DetectorId** | 0 | 0 | TRUE |
| **AlertTitle** | 0 | 0 | TRUE |
| **Category** | 0 | 0 | TRUE |
| **MitreTechniques** | 2733136 | 57.43791 | TRUE |
| **IncidentGrade** | 25775 | 0.541672 | TRUE |
| **ActionGrouped** | 4730205 | 99.40709 | TRUE |
| **ActionGranular** | 4730205 | 99.40709 | TRUE |
| **EntityType** | 0 | 0 | TRUE |
| **EvidenceRole** | 0 | 0 | TRUE |
| **DeviceId** | 0 | 0 | TRUE |
| **Sha256** | 0 | 0 | TRUE |
| **IpAddress** | 0 | 0 | TRUE |
| **Url** | 0 | 0 | TRUE |
| **AccountSid** | 0 | 0 | TRUE |
| **AccountUpn** | 0 | 0 | TRUE |
| **AccountObjectId** | 0 | 0 | TRUE |
| **AccountName** | 0 | 0 | TRUE |
| **DeviceName** | 0 | 0 | TRUE |
| **NetworkMessageId** | 0 | 0 | TRUE |
| **EmailClusterId** | 4710241 | 98.98754 | TRUE |
| **RegistryKey** | 0 | 0 | TRUE |
| **RegistryValueName** | 0 | 0 | TRUE |
| **RegistryValueData** | 0 | 0 | TRUE |
| **ApplicationId** | 0 | 0 | TRUE |
| **ApplicationName** | 0 | 0 | TRUE |
| **OAuthApplicationId** | 0 | 0 | TRUE |
| **ThreatFamily** | 4721038 | 99.21444 | TRUE |
| **FileName** | 0 | 0 | TRUE |
| **FolderPath** | 0 | 0 | TRUE |
| **ResourceIdName** | 0 | 0 | TRUE |
| **ResourceType** | 4754863 | 99.92529 | TRUE |
| **Roles** | 4649117 | 97.703 | TRUE |
| **OSFamily** | 0 | 0 | TRUE |
| **OSVersion** | 0 | 0 | TRUE |
| **AntispamDirection** | 4669846 | 98.13863 | TRUE |
| **SuspicionLevel** | 4037145 | 84.84217 | TRUE |
| **LastVerdict** | 3642221 | 76.54269 | TRUE |
| **CountryCode** | 0 | 0 | TRUE |
| **State** | 0 | 0 | TRUE |
| **City** | 0 | 0 | TRUE |

# Correlation results

## ANOVA Results for Numerical Features:

Unnamed: 0: F-statistic = 0.22925829642951948, p-value = 0.7951231382876303

Id: F-statistic = 1439.484198513051, p-value = 0.0

OrgId: F-statistic = 58997.76021886329, p-value = 0.0

IncidentId: F-statistic = 76452.55955725766, p-value = 0.0

AlertId: F-statistic = 47081.29268933218, p-value = 0.0

DetectorId: F-statistic = 32908.98954426741, p-value = 0.0

AlertTitle: F-statistic = 48290.78906666099, p-value = 0.0

DeviceId: F-statistic = 15635.311122018205, p-value = 0.0

Sha256: F-statistic = 47594.834053318205, p-value = 0.0

IpAddress: F-statistic = 58524.64720546779, p-value = 0.0

Url: F-statistic = 21112.06672260003, p-value = 0.0

AccountSid: F-statistic = 24636.7169202503, p-value = 0.0

AccountUpn: F-statistic = 4759.04937297763, p-value = 0.0

AccountObjectId: F-statistic = 26743.9579418214, p-value = 0.0

AccountName: F-statistic = 17730.550232143647, p-value = 0.0

DeviceName: F-statistic = 50484.52016148558, p-value = 0.0

NetworkMessageId: F-statistic = 63803.88275238493, p-value = 0.0

RegistryKey: F-statistic = 8767.964467241332, p-value = 0.0

RegistryValueName: F-statistic = 217.63178162648973, p-value = 3.076551888558765e-95

RegistryValueData: F-statistic = 404.24902410506155, p-value = 2.8305701665374938e-176

ApplicationId: F-statistic = 1467.8178662128046, p-value = 0.0

ApplicationName: F-statistic = 1440.178631903879, p-value = 0.0

OAuthApplicationId: F-statistic = 105.93472762386102, p-value = 9.866477532828021e-47

FileName: F-statistic = 62598.092000631914, p-value = 0.0

FolderPath: F-statistic = 63173.555928302405, p-value = 0.0

ResourceIdName: F-statistic = 863.3021208386185, p-value = 0.0

OSFamily: F-statistic = 10039.212454464861, p-value = 0.0

OSVersion: F-statistic = 10061.126800280788, p-value = 0.0

CountryCode: F-statistic = 228965.7741272291, p-value = 0.0

State: F-statistic = 188291.76701489132, p-value = 0.0

City: F-statistic = 190443.64795173952, p-value = 0.0

target\_encoded: F-statistic = inf, p-value = 0.0

## Chi-Square Results for Categorical Features:

Category: Chi2-statistic = 891259.7707364529, p-value = 0.0

IncidentGrade: Chi2-statistic = 9465285.999999998, p-value = 0.0

EntityType: Chi2-statistic = 895389.8854232655, p-value = 0.0

EvidenceRole: Chi2-statistic = 121913.79393155336, p-value = 0.0

# Feature Importance

22 target\_encoded 0.707158

0 OrgId 0.065122

1 IncidentId 0.055562

3 DetectorId 0.043640

4 AlertTitle 0.036899

2 AlertId 0.025943

5 Category 0.021037

19 CountryCode 0.008315

6 EntityType 0.006283

10 IpAddress 0.003457

21 City 0.003440

13 AccountObjectId 0.003271

15 DeviceName 0.003025

12 AccountSid 0.002987

16 NetworkMessageId 0.002952

14 AccountName 0.002581

20 State 0.002202

18 FolderPath 0.001450

7 EvidenceRole 0.001424

9 Sha256 0.001001

17 FileName 0.000948

11 Url 0.000669

8 DeviceId 0.000633

# Output With 1Lac Records

Accuracy on Test Data: 87.54%

Classification Report:

precision recall f1-score support

0 0.87 0.92 0.89 42352

1 0.84 0.78 0.81 21612

2 0.91 0.88 0.90 36036

accuracy 0.88 100000

macro avg 0.87 0.86 0.87 100000

weighted avg 0.88 0.88 0.87 100000

Confusion Matrix:

[[38789 1746 1817]

[ 3258 16897 1457]

[ 2755 1425 31856]]

# Output with 5lac records

Accuracy on Test Data: 90.39%

Classification Report:

precision recall f1-score support

0 0.89 0.94 0.92 211593

1 0.87 0.82 0.84 108338

2 0.94 0.91 0.93 180069

accuracy 0.90 500000

macro avg 0.90 0.89 0.89 500000

weighted avg 0.90 0.90 0.90 500000

Confusion Matrix:

[[198473 6981 6139]

[ 14355 88798 5185]

[ 9117 6262 164690]]

# Output with 10lac records

Accuracy on Test Data: 91.29%

Classification Report:

precision recall f1-score support

0 0.91 0.94 0.92 423011

1 0.88 0.83 0.85 217660

2 0.94 0.93 0.93 359329

accuracy 0.91 1000000

macro avg 0.91 0.90 0.90 1000000

weighted avg 0.91 0.91 0.91 100000

Confusion Matrix:

[[399224 13282 10505]

[ 26580 181138 9942]

[ 14685 12116 332528]]

# Evaluation with entire dataset

