Data Cleaning:

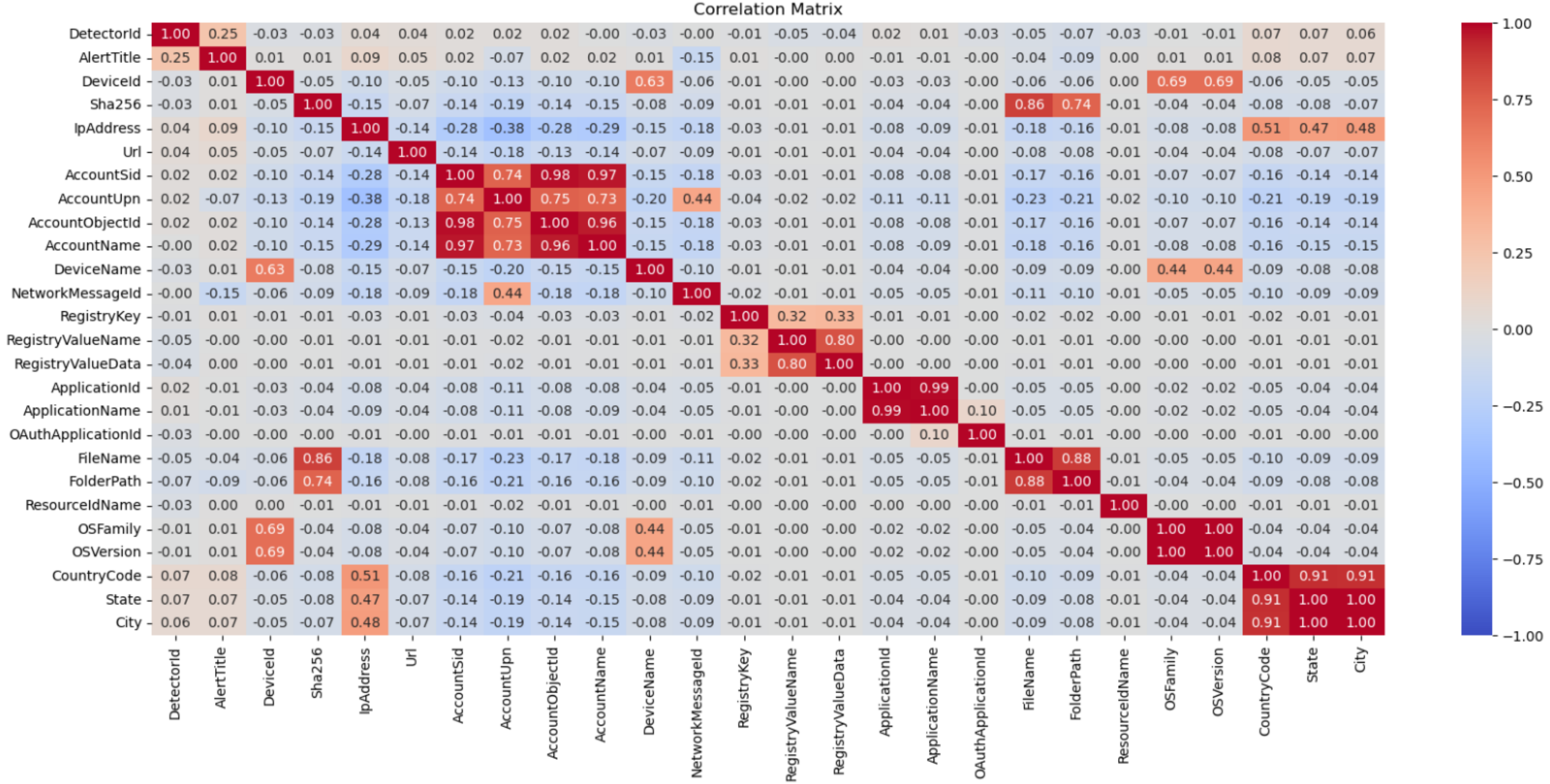
- To select the required columns read only initial 10000 rows to understand the importance of the columns.
- Initial ID related columns like ['Id', 'OrgId', 'IncidentId', 'AlertId'] can be excluded as they will not be adding any value to the model.
- To exclude imbalance in the data that we feed to the model have read column by column from the GUIDE_train data and calculated the number of columns with % blank as 40, 50, 60, 70, 80, 90 respectively
- Result

% cells blank	No of Columns
90% cells Blank	7
80% cells Blank	8
70% cells Blank	9
60% cells Blank	9
50% cells Blank	10
40% cells Blank	10

- Up on analysis it has been decided to drop the columns with 50% or more blank cells which are 10 in number and the columns are ['MitreTechniques', 'ActionGrouped', 'ActionGranular', 'EmailClusterId','ThreatFamily', 'ResourceType', 'Roles', 'AntispamDirection', 'SuspicionLevel', 'LastVerdict'].
- So, the finalised columns are ['Timestamp', 'DetectorId', 'AlertTitle', 'Category', 'IncidentGrade', 'EntityType', 'EvidenceRole', 'DeviceId', 'Sha256', 'IpAddress', 'Url', 'AccountSid', 'AccountUpn', 'AccountObjectId', 'AccountName', 'DeviceName', 'NetworkMessageId', 'RegistryKey', 'RegistryValueName', 'RegistryValueData', 'ApplicationId', 'ApplicationName', 'OAuthApplicationId', 'FileName', 'FolderPath', 'ResourceIdName', 'OSFamily', 'OSVersion', 'CountryCode', 'State', 'City'].
- After dropping the unnecessary columns, the rows with more than 20 blanks are being dropped.
- For achieving this we read the large file in chunks of 1,00,000 rows and then the rows with more than 20 blanks are being dropped and then added to the output file.
- And the output file is saved as 'GUIDE_train_cleaned.parquet' as this big file cannot be handled using CSV we are saving the output as .parquet file, achieved the file reduction by 75%

Feature Engineering:

- To balance the class imbalances we have made the stratification in such a way that the final output has equal number of TP, FP and BP which is 19,57,726 with the total size of the df as 58,73,178 rows.
- The balanced file is being saved as 'GUIDE_train_cleaned_balanced.parquet'.
- The balanced file has been split in to the rows of 5,00,000 and then saved as GUIDE_train_cleaned_0.parquet to GUIDE_train_cleaned_11.parquet files, for the purpose of Bigdata handling in future.
- Constructed correlation matrix for the existing columns as shown below



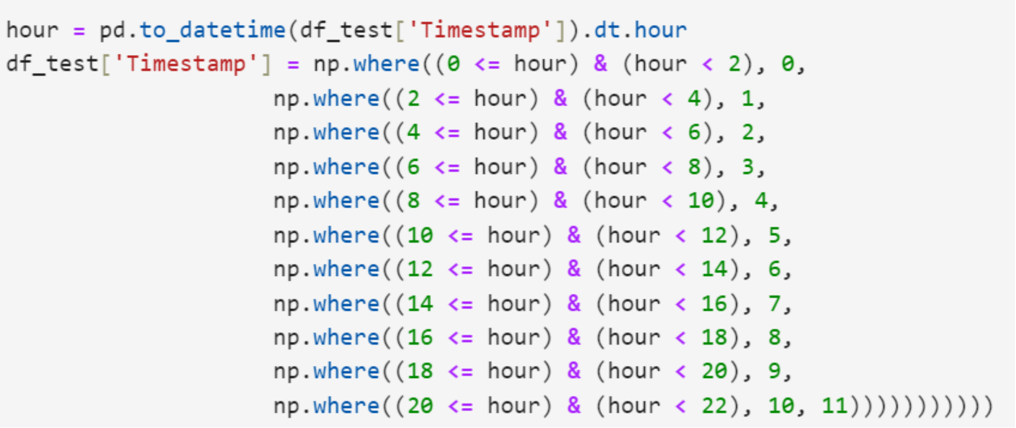
	DetectorId	AlertTitle	DeviceId	Sha256	IpAddress	Url	AccountSid	AccountUpn	AccountObjectId	AccountName	DeviceName	NetworkMessageId	RegistryKey	RegistryValueName	RegistryValueData	ApplicationId	ApplicationName	OAuthApplicationId	FileName	FolderPath	ResourceIdName	OSFamily	OSVersion	CountryCode	State	City
DetectorId	1.00	0.25	-0.03	-0.03	0.04	0.04	0.02	0.02	0.02	-0.00	-0.03	-0.00	-0.01	-0.05	-0.04	0.02	0.01	-0.03	-0.05	-0.07	-0.03	-0.01	-0.01	0.07	0.07	0.06
AlertTitle	0.25	1.00	0.01	0.01	0.09	0.05	0.02	-0.07	0.02	0.02	0.01	-0.15	0.01	-0.00	0.00	-0.01	-0.01	-0.00	-0.04	-0.09	0.00	0.01	0.01	0.08	0.07	0.07
DeviceId	-0.03	0.01	1.00	-0.05	-0.10	-0.05	-0.10	-0.13	-0.10	-0.10	0.63	-0.06	-0.01	-0.00	-0.00	-0.03	-0.03	-0.00	-0.06	-0.06	0.00	0.69	0.69	-0.06	-0.05	-0.05
Sha256	-0.03	0.01	-0.05	1.00	-0.15	-0.07	-0.14	-0.19	-0.14	-0.15	-0.08	-0.09	-0.01	-0.01	-0.01	-0.04	-0.04	-0.00	0.86	0.74	-0.01	-0.04	-0.04	-0.08	-0.08	-0.07
IpAddress	0.04	0.09	-0.10	-0.15	1.00	-0.14	-0.28	-0.38	-0.28	-0.29	-0.15	-0.18	-0.03	-0.01	-0.01	-0.08	-0.09	-0.01	-0.18	-0.16	-0.01	-0.08	-0.08	0.51	0.47	0.48
Url	0.04	0.05	-0.05	-0.07	-0.14	1.00	-0.14	-0.18	-0.13	-0.14	-0.07	-0.09	-0.01	-0.01	-0.01	-0.04	-0.04	-0.00	-0.08	-0.08	-0.01	-0.04	-0.04	-0.08	-0.07	-0.07
AccountSid	0.02	0.02	-0.10	-0.14	-0.28	-0.14	1.00	0.74	0.98	0.97	-0.15	-0.18	-0.03	-0.01	-0.01	-0.08	-0.08	-0.01	-0.17	-0.16	-0.01	-0.07	-0.07	-0.16	-0.14	-0.14
AccountUpn	0.02	-0.07	-0.13	-0.19	-0.38	-0.18	0.74	1.00	0.75	0.73	-0.20	0.44	-0.04	-0.02	-0.02	-0.11	-0.11	-0.01	-0.23	-0.21	-0.02	-0.10	-0.10	-0.21	-0.19	-0.19
AccountObjectId	0.02	0.02	-0.10	-0.14	-0.28	-0.13	0.98	0.75	1.00	0.96	-0.15	-0.18	-0.03	-0.01	-0.01	-0.08	-0.08	-0.01	-0.17	-0.16	-0.01	-0.07	-0.07	-0.16	-0.14	-0.14
AccountName	-0.00	0.02	-0.10	-0.15	-0.29	-0.14	0.97	0.73	0.96	1.00	-0.15	-0.18	-0.03	-0.01	-0.01	-0.08	-0.09	-0.01	-0.18	-0.16	-0.01	-0.08	-0.08	-0.16	-0.15	-0.15
DeviceName	-0.03	0.01	0.63	-0.08	-0.15	-0.07	-0.15	-0.20	-0.15	-0.15	1.00	-0.10	-0.01	-0.01	-0.01	-0.04	-0.04	-0.00	-0.09	-0.09	-0.00	0.44	0.44	-0.09	-0.08	-0.08
NetworkMessageId	-0.00	-0.15	-0.06	-0.09	-0.18	-0.09	-0.18	0.44	-0.18	-0.18	-0.10	1.00	-0.02	-0.01	-0.01	-0.05	-0.05	-0.01	-0.11	-0.10	-0.01	-0.05	-0.05	-0.10	-0.09	-0.09
RegistryKey	-0.01	0.01	-0.01	-0.01	-0.03	-0.01	-0.03	-0.04	-0.03	-0.03	-0.01	-0.02	1.00	0.32	0.33	-0.01	-0.01	-0.00	-0.02	-0.02	-0.00	-0.01	-0.01	-0.02	-0.01	-0.01
RegistryValueName	-0.05	-0.00	-0.00	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	0.32	1.00	0.80	-0.00	-0.00	-0.00	-0.01	-0.01	-0.00	-0.00	-0.00	-0.01	-0.01	-0.01
RegistryValueData	-0.04	0.00	-0.00	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	0.33	0.80	1.00	-0.00	-0.00	-0.00	-0.01	-0.01	-0.00	-0.00	-0.00	-0.01	-0.01	-0.01
ApplicationId	0.02	-0.01	-0.03	-0.04	-0.08	-0.04	-0.08	-0.11	-0.08	-0.08	-0.04	-0.05	-0.01	-0.00	-0.00	1.00	0.99	-0.00	-0.05	-0.05	-0.00	-0.02	-0.02	-0.05	-0.04	-0.04
ApplicationName	0.01	-0.01	-0.03	-0.04	-0.09	-0.04	-0.08	-0.11	-0.08	-0.09	-0.04	-0.05	-0.01	-0.00	-0.00	0.99	1.00	0.10	-0.05	-0.05	-0.00	-0.02	-0.02	-0.05	-0.04	-0.04
OAuthApplicationId	-0.03	-0.00	-0.00	-0.00	-0.01	-0.00	-0.01	-0.01	-0.01	-0.01	-0.00	-0.01	-0.00	-0.00	-0.00	-0.00	0.10	1.00	-0.01	-0.01	-0.00	-0.00	-0.00	-0.01	-0.00	-0.00
FileName	-0.05	-0.04	-0.06	0.86	-0.18	-0.08	-0.17	-0.23	-0.17	-0.18	-0.09	-0.11	-0.02	-0.01	-0.01	-0.05	-0.05	-0.01	1.00	0.88	-0.01	-0.05	-0.05	-0.10	-0.09	-0.09
FolderPath	-0.07	-0.09	-0.06	0.74	-0.16	-0.08	-0.16	-0.21	-0.16	-0.16	-0.09	-0.10	-0.02	-0.01	-0.01	-0.05	-0.05	-0.01	0.88	1.00	-0.01	-0.04	-0.04	-0.09	-0.08	-0.08
ResourceIdName	-0.03	0.00	0.00	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	-0.01	-0.00	-0.01	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.01	-0.01	1.00	-0.00	-0.00	-0.01	-0.01	-0.01
OSFamily	-0.01	0.01	0.69	-0.04	-0.08	-0.04	-0.07	-0.10	-0.07	-0.08	0.44	-0.05	-0.01	-0.00	-0.00	-0.02	-0.02	-0.00	-0.05	-0.04	-0.00	1.00	1.00	-0.04	-0.04	-0.04
OSVersion	-0.01	0.01	0.69	-0.04	-0.08	-0.04	-0.07	-0.10	-0.07	-0.08	0.44	-0.05	-0.01	-0.00	-0.00	-0.02	-0.02	-0.00	-0.05	-0.04	-0.00	1.00	1.00	-0.04	-0.04	-0.04
CountryCode	0.07	0.08	-0.06	-0.08	0.51	-0.08	-0.16	-0.21	-0.16	-0.16	-0.09	-0.10	-0.02	-0.01	-0.01	-0.05	-0.05	-0.01	-0.10	-0.09	-0.01	-0.04	-0.04	1.00	0.91	0.91
State	0.07	0.07	-0.05	-0.08	0.47	-0.07	-0.14	-0.19	-0.14	-0.15	-0.08	-0.09	-0.01	-0.01	-0.01	-0.04	-0.04	-0.00	-0.09	-0.08	-0.01	-0.04	-0.04	0.91	1.00	1.00
City	0.06	0.07	-0.05	-0.07	0.48	-0.07	-0.14	-0.19	-0.14	-0.15	-0.08	-0.09	-0.01	-0.01	-0.01	-0.04	-0.04	-0.00	-0.09	-0.08	-0.01	-0.04	-0.04	0.91	1.00	1.00

The highest correlated pairs are as below

Feature 1	Feature 2	Correlation
OSFamily	OSVersion	0.999257
State	City	0.995713
ApplicationId	ApplicationName	0.987270
AccountObjectId	AccountSid	0.983536
AccountSid	AccountName	0.966836
AccountObjectId	AccountName	0.958557
State	CountryCode	0.914400
City	CountryCode	0.910737
FolderPath	FileName	0.876606
FileName	Sha256	0.859692
RegistryValueName	RegistryValueData	0.802317
AccountUpn	AccountObjectId	0.749224
AccountSid	AccountUpn	0.739470
FolderPath	Sha256	0.739084
AccountName	AccountUpn	0.729920
OSVersion	DeviceId	0.688451
OSFamily	DeviceId	0.687887
DeviceName	DeviceId	0.633694
CountryCode	IpAddress	0.508536

- And the following actions are bding taken for the same
- Drop OSFamily (Retain OSVersion)
  - Reasoning: OSFamily and OSVersion are highly correlated (0.999257). OSVersion provides more detailed information, so it’s preferable to keep it over OSFamily.
- Drop City (Retain State)
  - Reasoning: City and State are highly correlated (0.995713). State represents a broader geographic level, which is often more useful for analysis.
- Drop ApplicationId (Retain ApplicationName)
  - Reasoning: ApplicationId and ApplicationName are highly correlated (0.987270). ApplicationName is more interpretable and user-friendly, making it the better choice to retain.
- Drop AccountObjectId (Retain AccountSid)
  - Reasoning: AccountObjectId and AccountSid are highly correlated (0.983536). AccountSid is typically more directly associated with user accounts, making it more relevant for most analyses.
- Keep both AccountName and AccountSid
  - Reasoning: While AccountSid and AccountName are highly correlated (0.966836), they provide distinct and potentially valuable information about user accounts, so both are retained.

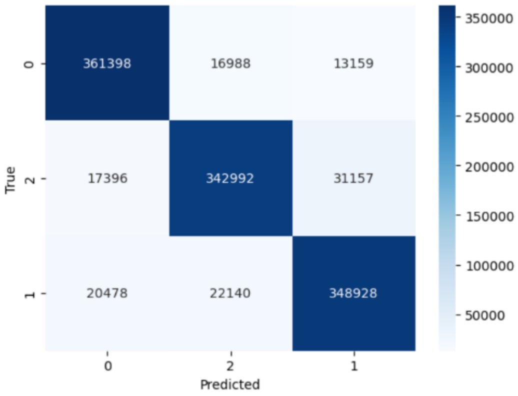
- Drop CountryCode (Retain State)
  - Reasoning: CountryCode is highly correlated with both State (0.914400) and City (0.910737). State is more granular than CountryCode, making it more useful for detailed geographic analysis.
- Combine FileName and FolderPath into FullFilePath
  - Reasoning: FolderPath and FileName are correlated (0.876606). Combining them into a single FullFilePath column simplifies the data structure and provides a comprehensive path identifier for file-related analysis.
- Keep both RegistryValueName and RegistryValueData
  - Reasoning: Although they are correlated (0.802317), RegistryValueName and RegistryValueData together offer crucial details about registry entries, which can be important in various analyses.
- Drop FileName (Already handled in Step 7)
  - Reasoning: FileName was already combined with FolderPath in Step 7, so it's no longer needed as a separate column.
- Keep both DeviceId and DeviceName
  - Reasoning: DeviceId and DeviceName have a moderate correlation (0.633694). Both provide unique identifiers for devices that may be useful in different contexts, so they are both retained.
- Drop AccountObjectId (Already handled in Step 4)
  - Reasoning: AccountObjectId was already dropped in Step 4, so no further action is necessary.
- Convert the Timestamp column into hour group basis the below logic



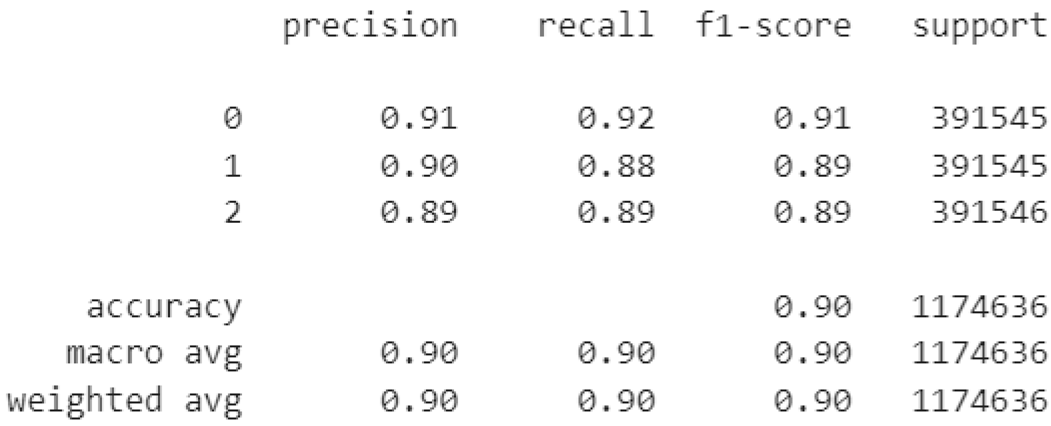
```
hour = pd.to_datetime(df_test['Timestamp']).dt.hour
df_test['Timestamp'] = np.where((0 <= hour) & (hour < 2), 0,
                       np.where((2 <= hour) & (hour < 4), 1,
                       np.where((4 <= hour) & (hour < 6), 2,
                       np.where((6 <= hour) & (hour < 8), 3,
                       np.where((8 <= hour) & (hour < 10), 4,
                       np.where((10 <= hour) & (hour < 12), 5,
                       np.where((12 <= hour) & (hour < 14), 6,
                       np.where((14 <= hour) & (hour < 16), 7,
                       np.where((16 <= hour) & (hour < 18), 8,
                       np.where((18 <= hour) & (hour < 20), 9,
                       np.where((20 <= hour) & (hour < 22), 10, 11)))))))))))
```

- This resultant dataframe will look like (5873178, 77).

Baseline Model:



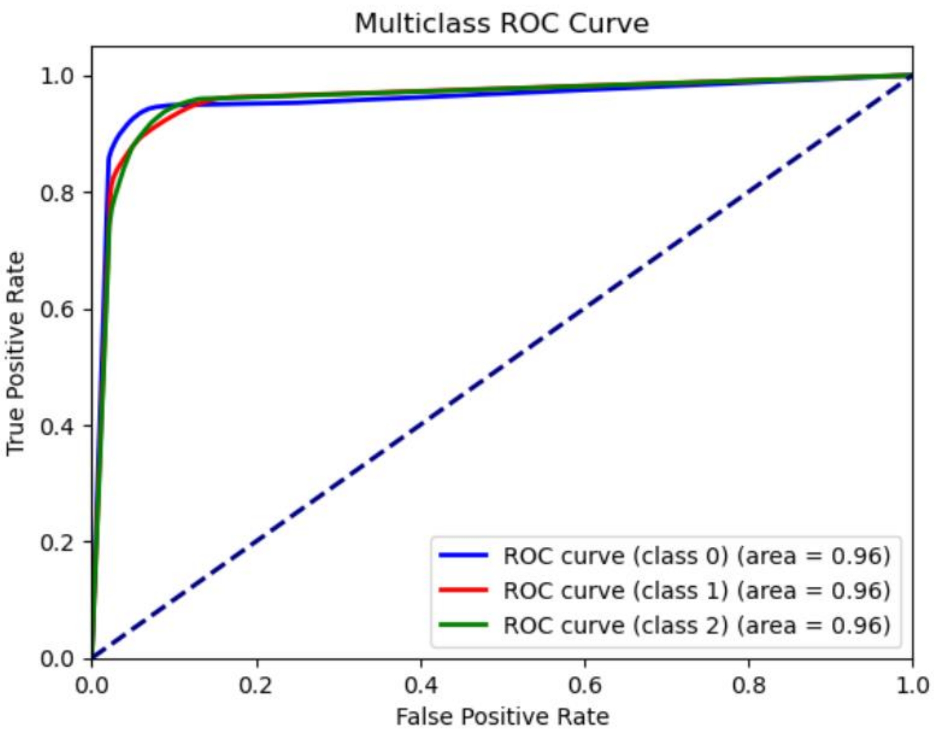
- For the purpose of Baseline model Decision Tree classifier has been choosen, by splitting the data in to 80 20 for training and testing.
- which has resulted in a accuracy score of 89% and the confusion matrix and ROC curves for the same is as shown below

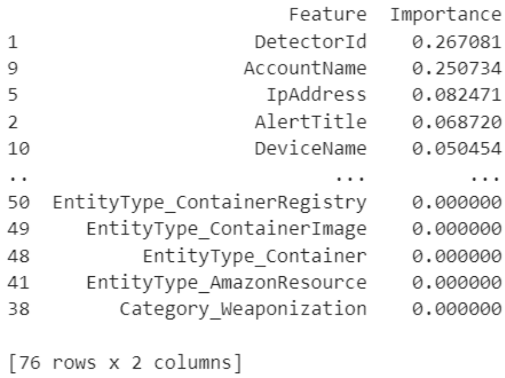


```
              precision    recall  f1-score   support

           0       0.91      0.92      0.91    391545
           1       0.90      0.88      0.89    391545
           2       0.89      0.89      0.89    391546

    accuracy                           0.90   1174636
   macro avg       0.90      0.90      0.90   1174636
weighted avg       0.90      0.90      0.90   1174636
```





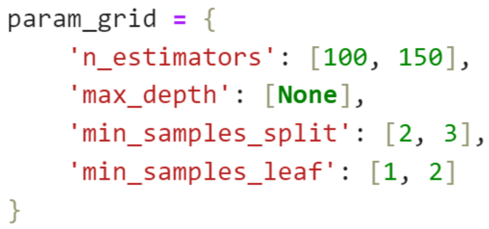
```
                       Feature  Importance
1                   DetectorId    0.267081
9                  AccountName    0.250734
5                    IpAddress    0.082471
2                   AlertTitle    0.068720
10                  DeviceName    0.050454
..                         ...         ...
50  EntityType_ContainerRegistry    0.000000
49     EntityType_ContainerImage    0.000000
48          EntityType_Container    0.000000
41     EntityType_AmazonResource    0.000000
38        Category_Weaponization    0.000000

[76 rows x 2 columns]
```

- The feature importance for the columns are as show above
- The cross validation score is: [0.89750272 0.89721497 0.89722774 0.89712634 0.89748135] with Mean CV score: 0.897310621266109.
- This is a strong accuracy score, indicating that the model is correctly classifying about 90% of the samples.
- Precision: Measures the accuracy of the positive predictions. High values for all classes indicate that when the model predicts a class, it is often correct.
- Measures how well the model captures all positive cases of each class. High values suggest that the model is good at identifying instances of each class.
- The harmonic mean of precision and recall. High values indicate a good balance between precision and recall.
- The high ROC AUC scores (close to 1) for all classes suggest excellent performance in distinguishing between the classes.
- Consistent crossvalidation scores reinforce the model's robustness and generalizability across different data splits.
- DetectorId, AccountName, IpAddress are the most important features, which makes sense if they have a significant impact on classification.

Grid Search:

- As the results for the decision tree are good, we will be proceeding with random forest further.
- To find the best parameters set for Random Forest we will be doing the grid search with the ranges as mentioned below.

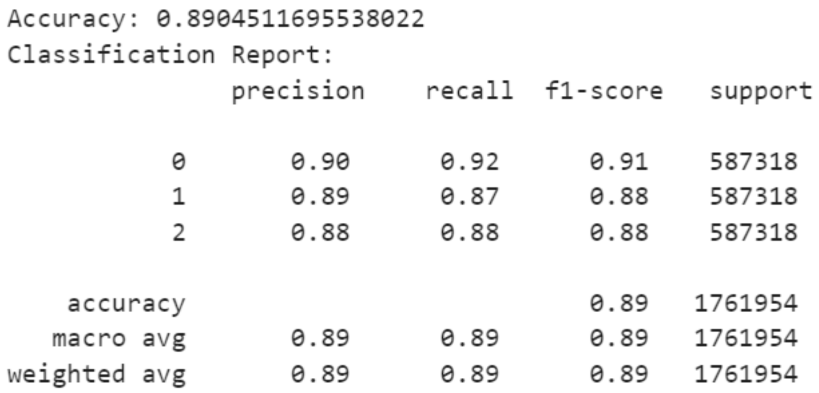


```
param_grid = {
    'n_estimators': [100, 150],
    'max_depth': [None],
    'min_samples_split': [2, 3],
    'min_samples_leaf': [1, 2]
}
```

- The best parameters was found to be Best Parameters: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}.

Final Model:

- By splitting the data into 70 30 we will train the model with the best patameters that we have attained from grid search.
- The output and the classification report are as shown below on train data

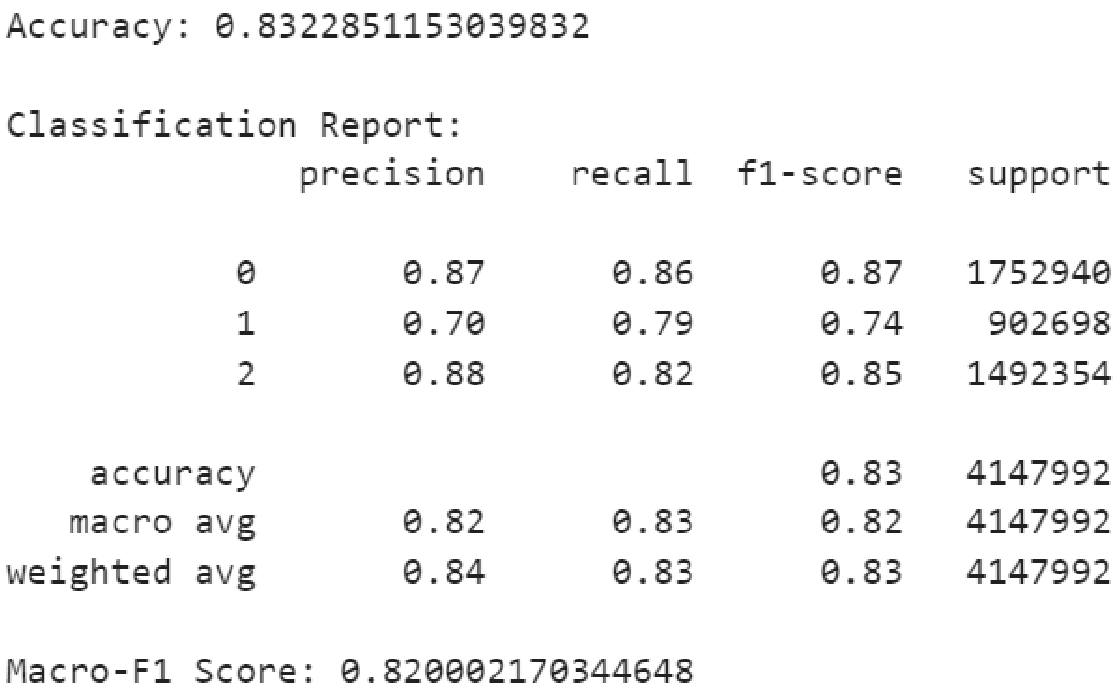


```
Accuracy: 0.8904511695538022
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.92      0.91    587318
           1       0.89      0.87      0.88    587318
           2       0.88      0.88      0.88    587318

    accuracy                           0.89   1761954
   macro avg       0.89      0.89      0.89   1761954
weighted avg       0.89      0.89      0.89   1761954
```

- Doing the preprocessing and feature engineering for the test data the outputs are as shown below



```
Accuracy: 0.8322851153039832

Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.86      0.87   1752940
           1       0.70      0.79      0.74    902698
           2       0.88      0.82      0.85   1492354

    accuracy                           0.83   4147992
   macro avg       0.82      0.83      0.82   4147992
weighted avg       0.84      0.83      0.83   4147992

Macro-F1 Score: 0.820002170344648
```