/PySpark/Py_spark...

Modeling on Malware Data Using PySpark.

READY

Importing required libraries

READY

```
%pyspark
from pyspark.sql import SparkSession
#sqlContext = SQLContext(sc)
```

READY

Reading Malware data from HDFS

READY

```
%pyspark
data = spark.read.csv('hdfs://172.27.35.78:9000/MicrosoftMalware.csv',header=True, inferSchema=True)
```

READY

View data using “show” command

READY

```
%pyspark
data.show(2,True)
```

READY

```
+---+--------------------+------------+-------------+---------------+------------+------+----------------+---
-------------+-------------------------+-------------------+-----------------+------+-----------------+------
--------+----------------------+-----------------+---------------------------+---------+---------+--------+--
-----+-------+--------------------+--------------------+----------+-----------+---------------+-----+--------
-------+------------+--------+------------+---------------------+-------------------+------------------------
+-------------------------+-------------------------+--------------------------------------+-----------------
--------------+-------------------------------+--------------------------+--------------------------------+--
------------------------+-----------------------+----------------------+-------------------------------------
------------+-------------------------------------------------+----------------------------------------------
-+----------------------------+-------------------------------------+----------------+---------------------+-
--------------+--------------------+----------------------+----------------+----------------+----------------
--------+----------------------------------+---------------------------+--------------------------------+----
----------------------------+-----------------------+------------------------+------------------------+------
-----------+-------------------------------------+--------------------------------+--------------------------
+----------------------+---------------------+-------------------+---------------------------------------+---
---------+---------------------+-------------+
|_c0|   MachineIdentifier| ProductName|EngineVersion|     AppVersion|AvSigVersion|IsBeta|RtpStateBitfield|IsS
xsPassiveMode|AVProductStatesIdentifier|AVProductsInstalled|AVProductsEnabled|HasTpm|CountryIdentifier|CityId
entifier|OrganizationIdentifier|GeoNameIdentifier|LocaleEnglishNameIdentifier| Platform|Processor|   OsVer|Os
Build|OsSuite|OsPlatformSubRelease|          OsBuildLab|SkuEdition|IsProtected|AutoSampleOptIn|SMode|IeVerIde
ntifier| SmartScreen|Firewall|UacLuaenable|Census_MDC2FormFactor|Census_DeviceFamily|Census_OEMNameIdentifier
|Census_OEMModelIdentifier|Census_ProcessorCoreCount|Census_ProcessorManufacturerIdentifier|Census_ProcessorM
odelIdentifier|Census_PrimaryDiskTotalCapacity|Census_PrimaryDiskTypeName|Census_SystemVolumeTotalCapacity|Ce
nsus_HasOpticalDiskDrive|Census_TotalPhysicalRAM|Census_ChassisTypeName|Census_InternalPrimaryDiagonalDisplay
SizeInInches|Census_InternalPrimaryDisplayResolutionHorizontal|Census_InternalPrimaryDisplayResolutionVertica
l|Census_PowerPlatformRoleName|Census_InternalBatteryNumberOfCharges|Census_OSVersion|Census_OSArchitecture|C
ensus_OSBranch|Census_OSBuildNumber|Census_OSBuildRevision|Census_OSEdition|Census_OSSkuName|Census_OSInstall
```

```
TypeName|Census_OSInstallLanguageIdentifier|Census_OSUILocaleIdentifier|Census_OSWUAutoUpdateOptionsName|Cens
us_IsPortableOperatingSystem|Census_GenuineStateName|Census_ActivationChannel|Census_IsFlightsDisabled|Census
_FlightRing|Census_FirmwareManufacturerIdentifier|Census_FirmwareVersionIdentifier|Census_IsSecureBootEnabled
|Census_IsVirtualDevice|Census_IsTouchEnabled|Census_IsPenCapable|Census_IsAlwaysOnAlwaysConnectedCapable|Wdf
t_IsGamer|Wdft_RegionIdentifier|HasDetections|
+---+--------------------+------------+-------------+---------------+------------+------+----------------+---
-------------+-------------------------+-------------------+-----------------+------+-----------------+------
--------+----------------------+-----------------+---------------------------+---------+---------+--------+--
-----+-------+--------------------+--------------------+----------+-----------+---------------+-----+--------
-------+------------+--------+------------+---------------------+-------------------+------------------------
+-------------------------+-------------------------+--------------------------------------+-----------------
--------------+-------------------------------+--------------------------+--------------------------------+--
------------------------+-----------------------+----------------------+-------------------------------------
------------+-------------------------------------------------+----------------------------------------------
-+----------------------------+-------------------------------------+----------------+---------------------+-
--------------+--------------------+----------------------+----------------+----------------+----------------
--------+----------------------------------+---------------------------+--------------------------------+----
----------------------------+-----------------------+------------------------+------------------------+------
-----------+-------------------------------------+--------------------------------+--------------------------
+----------------------+---------------------+-------------------+---------------------------------------+---
---------+---------------------+-------------+
|  2|000007905a28d863f...|win8defender|  1.1.15100.1|4.18.1807.18075|1.273.1341.0|     0|             7.0|
0|                  53447.0|                1.0|              1.0|     1|               86|      153579.0|
18.0|             64.0|                         49|windows10|      x64|10.0.0.0|  17134|    768|
rs4|17134.1.amd64fre....|      Home|        1.0|              0|  0.0|          137.0|RequireAdmin|     1.0|
1.0|              Desktop|    Windows.Desktop|                  4908.0|                 317701.0|
4.0|                                   5.0|                         1972.0|                       114473.0|
SSD|                        113907.0|                         0|                 4096.0|               Deskto
p|                                             21.5|                                           1920.0|
1080.0|                     Desktop|                          4.2949673E9|  10.0.17134.165|                am
d64|    rs4_release|               17134|                   165|            Core|            CORE|
UUPUpgrade|                               7.0|                         30|                        FullAuto|
0|             IS_GENUINE|              OEM:NONSLP|                     0.0|           Retail|
142.0|                         52682.0|                         0|                   0.0|
0|                  0|                                    0.0|         0.0|                  3.0|
0|
|  5|000016191b897145d...|win8defender|  1.1.15100.1|4.18.1807.18075|1.273.1094.0|     0|             7.0|
0|                  53447.0|                1.0|              1.0|     1|               97|       13598.0|
27.0|            126.0|                        124|windows10|      x64|10.0.0.0|  17134|    256|
rs4|17134.1.amd64fre....|       Pro|        1.0|              0|  0.0|          137.0|RequireAdmin|     1.0|
1.0|              Desktop|    Windows.Desktop|                  3800.0|                 340727.0|
2.0|                                   5.0|                         4324.0|                       114473.0|
SSD|                        113671.0|                         0|                 8192.0|               Deskto
p|                                             21.5|                                           1920.0|
1080.0|                     Desktop|                          4.2949673E9|  10.0.17134.165|                am
d64|    rs4_release|               17134|                   165|    Professional|    PROFESSIONAL|
UUPUpgrade|                              18.0|                         72|                        FullAuto|
0|             IS_GENUINE|                  Retail|                     0.0|           Retail|
93.0|                         51039.0|                         0|                   0.0|                    0
|                  0|                                    0.0|         0.0|                 15.0|            1
|
+---+--------------------+------------+-------------+---------------+------------+------+----------------+---
-------------+-------------------------+-------------------+-----------------+------+-----------------+------
--------+----------------------+-----------------+---------------------------+---------+---------+--------+--
-----+-------+--------------------+--------------------+----------+-----------+---------------+-----+--------
-------+------------+--------+------------+---------------------+-------------------+------------------------
+-------------------------+-------------------------+--------------------------------------+-----------------
--------------+-------------------------------+--------------------------+--------------------------------+--
------------------------+-----------------------+----------------------+-------------------------------------
------------+-------------------------------------------------+----------------------------------------------
-+----------------------------+-------------------------------------+----------------+---------------------+-
```

```
--------------+--------------------+----------------------+----------------+----------------+---------------
--------+----------------------------------+---------------------------+----------------------------------+----
---------------------------+----------------------+-----------------------+-----------------------+------
-----------+------------------------------------+-------------------------------+-------------------------
+----------------------+--------------------+-------------------+----------------------------------------+---
---------+---------------------+-------------+
only showing top 2 rows
```

Printing the Schema of the features

READY

```
%pyspark
data.printSchema()
```

READY

```
root
 |-- _c0: integer (nullable = true)
 |-- MachineIdentifier: string (nullable = true)
 |-- ProductName: string (nullable = true)
 |-- EngineVersion: string (nullable = true)
 |-- AppVersion: string (nullable = true)
 |-- AvSigVersion: string (nullable = true)
 |-- IsBeta: integer (nullable = true)
 |-- RtpStateBitfield: double (nullable = true)
 |-- IsSxsPassiveMode: integer (nullable = true)
 |-- AVProductStatesIdentifier: double (nullable = true)
 |-- AVProductsInstalled: double (nullable = true)
 |-- AVProductsEnabled: double (nullable = true)
 |-- HasTpm: integer (nullable = true)
 |-- CountryIdentifier: integer (nullable = true)
 |-- CityIdentifier: double (nullable = true)
 |-- OrganizationIdentifier: double (nullable = true)
 |-- GeoNameIdentifier: double (nullable = true)
 |-- LocaleEnglishNameIdentifier: integer (nullable = true)
 |-- Platform: string (nullable = true)
 |-- Processor: string (nullable = true)
 |-- OsVer: string (nullable = true)
 |-- OsBuild: integer (nullable = true)
 |-- OsSuite: integer (nullable = true)
 |-- OsPlatformSubRelease: string (nullable = true)
 |-- OsBuildLab: string (nullable = true)
 |-- SkuEdition: string (nullable = true)
 |-- IsProtected: double (nullable = true)
 |-- AutoSampleOptIn: integer (nullable = true)
 |-- SMode: double (nullable = true)
 |-- IeVerIdentifier: double (nullable = true)
 |-- SmartScreen: string (nullable = true)
 |-- Firewall: double (nullable = true)
 |-- UacLuaenable: double (nullable = true)
 |-- Census_MDC2FormFactor: string (nullable = true)
 |-- Census_DeviceFamily: string (nullable = true)
 |-- Census_OEMNameIdentifier: double (nullable = true)
 |-- Census_OEMModelIdentifier: double (nullable = true)
 |-- Census_ProcessorCoreCount: double (nullable = true)
 |-- Census_ProcessorManufacturerIdentifier: double (nullable = true)
 |-- Census_ProcessorModelIdentifier: double (nullable = true)
 |-- Census_PrimaryDiskTotalCapacity: double (nullable = true)
 |-- Census_PrimaryDiskTypeName: string (nullable = true)
 |-- Census_SystemVolumeTotalCapacity: double (nullable = true)
 |-- Census_HasOpticalDiskDrive: integer (nullable = true)
 |-- Census_TotalPhysicalRAM: double (nullable = true)
 |-- Census_ChassisTypeName: string (nullable = true)
 |-- Census_InternalPrimaryDiagonalDisplaySizeInInches: double (nullable = true)
 |-- Census_InternalPrimaryDisplayResolutionHorizontal: double (nullable = true)
 |-- Census_InternalPrimaryDisplayResolutionVertical: double (nullable = true)
 |-- Census_PowerPlatformRoleName: string (nullable = true)
 |-- Census_InternalBatteryNumberOfCharges: double (nullable = true)
 |-- Census_OSVersion: string (nullable = true)
 |-- Census_OSArchitecture: string (nullable = true)
 |-- Census_OSBranch: string (nullable = true)
 |-- Census_OSBuildNumber: integer (nullable = true)
 |-- Census_OSBuildRevision: integer (nullable = true)
 |-- Census_OSEdition: string (nullable = true)
 |-- Census_OSSkuName: string (nullable = true)
 |-- Census_OSInstallTypeName: string (nullable = true)
 |-- Census_OSInstallLanguageIdentifier: double (nullable = true)
 |-- Census_OSUILocaleIdentifier: integer (nullable = true)
```

view the datatype of our DataFrame. READY

READY

```
%pyspark
type(data)
```

```
<class 'pyspark.sql.dataframe.DataFrame'>
```

Description of the data. READY

READY

```
%pyspark
data.describe()
```

```
DataFrame[summary: string, _c0: string, MachineIdentifier: string, ProductName: string, EngineVersion: string
, AppVersion: string, AvSigVersion: string, IsBeta: string, RtpStateBitfield: string, IsSxsPassiveMode: strin
g, AVProductStatesIdentifier: string, AVProductsInstalled: string, AVProductsEnabled: string, HasTpm: string,
CountryIdentifier: string, CityIdentifier: string, OrganizationIdentifier: string, GeoNameIdentifier: string,
LocaleEnglishNameIdentifier: string, Platform: string, Processor: string, OsVer: string, OsBuild: string, OsS
uite: string, OsPlatformSubRelease: string, OsBuildLab: string, SkuEdition: string, IsProtected: string, Auto
SampleOptIn: string, SMode: string, IeVerIdentifier: string, SmartScreen: string, Firewall: string, UacLuaena
ble: string, Census_MDC2FormFactor: string, Census_DeviceFamily: string, Census_OEMNameIdentifier: string, Ce
nsus_OEMModelIdentifier: string, Census_ProcessorCoreCount: string, Census_ProcessorManufacturerIdentifier: s
tring, Census_ProcessorModelIdentifier: string, Census_PrimaryDiskTotalCapacity: string, Census_PrimaryDiskTy
peName: string, Census_SystemVolumeTotalCapacity: string, Census_HasOpticalDiskDrive: string, Census_TotalPhy
sicalRAM: string, Census_ChassisTypeName: string, Census_InternalPrimaryDiagonalDisplaySizeInInches: string,
Census_InternalPrimaryDisplayResolutionHorizontal: string, Census_InternalPrimaryDisplayResolutionVertical: s
tring, Census_PowerPlatformRoleName: string, Census_InternalBatteryNumberOfCharges: string, Census_OSVersion:
string, Census_OSArchitecture: string, Census_OSBranch: string, Census_OSBuildNumber: string, Census_OSBuildR
evision: string, Census_OSEdition: string, Census_OSSkuName: string, Census_OSInstallTypeName: string, Census
_OSInstallLanguageIdentifier: string, Census_OSUILocaleIdentifier: string, Census_OSWUAutoUpdateOptionsName:
```

Get the columns of the DataFrame. READY

```
['_c0', 'MachineIdentifier', 'ProductName', 'EngineVersion', 'AppVersion', 'AvSigVersion', 'IsBeta', 'RtpStat
eBitfield', 'IsSxsPassiveMode', 'AVProductStatesIdentifier', 'AVProductsInstalled', 'AVProductsEnabled', 'Has
Tpm', 'CountryIdentifier', 'CityIdentifier', 'OrganizationIdentifier', 'GeoNameIdentifier', 'LocaleEnglishNam
eIdentifier', 'Platform', 'Processor', 'OsVer', 'OsBuild', 'OsSuite', 'OsPlatformSubRelease', 'OsBuildLab', '
SkuEdition', 'IsProtected', 'AutoSampleOptIn', 'SMode', 'IeVerIdentifier', 'SmartScreen', 'Firewall', 'UacLua
enable', 'Census_MDC2FormFactor', 'Census_DeviceFamily', 'Census_OEMNameIdentifier', 'Census_OEMModelIdentifi
er', 'Census_ProcessorCoreCount', 'Census_ProcessorManufacturerIdentifier', 'Census_ProcessorModelIdentifier'
, 'Census_PrimaryDiskTotalCapacity', 'Census_PrimaryDiskTypeName', 'Census_SystemVolumeTotalCapacity', 'Censu
s_HasOpticalDiskDrive', 'Census_TotalPhysicalRAM', 'Census_ChassisTypeName', 'Census_InternalPrimaryDiagonalD
isplaySizeInInches', 'Census_InternalPrimaryDisplayResolutionHorizontal', 'Census_InternalPrimaryDisplayResol
utionVertical', 'Census_PowerPlatformRoleName', 'Census_InternalBatteryNumberOfCharges', 'Census_OSVersion',
'Census_OSArchitecture', 'Census_OSBranch', 'Census_OSBuildNumber', 'Census_OSBuildRevision', 'Census_OSEditi
on', 'Census_OSSkuName', 'Census_OSInstallTypeName', 'Census_OSInstallLanguageIdentifier', 'Census_OSUILocale
Identifier', 'Census_OSWUAutoUpdateOptionsName', 'Census_IsPortableOperatingSystem', 'Census_GenuineStateName
', 'Census_ActivationChannel', 'Census_IsFlightsDisabled', 'Census_FlightRing', 'Census_FirmwareManufacturerI
dentifier', 'Census_FirmwareVersionIdentifier', 'Census_IsSecureBootEnabled', 'Census_IsVirtualDevice', 'Cens
us_IsTouchEnabled', 'Census_IsPenCapable', 'Census_IsAlwaysOnAlwaysConnectedCapable', 'Wdft_IsGamer', 'Wdft_R
```

Viewing the distribution of the target variable.

READY

READY

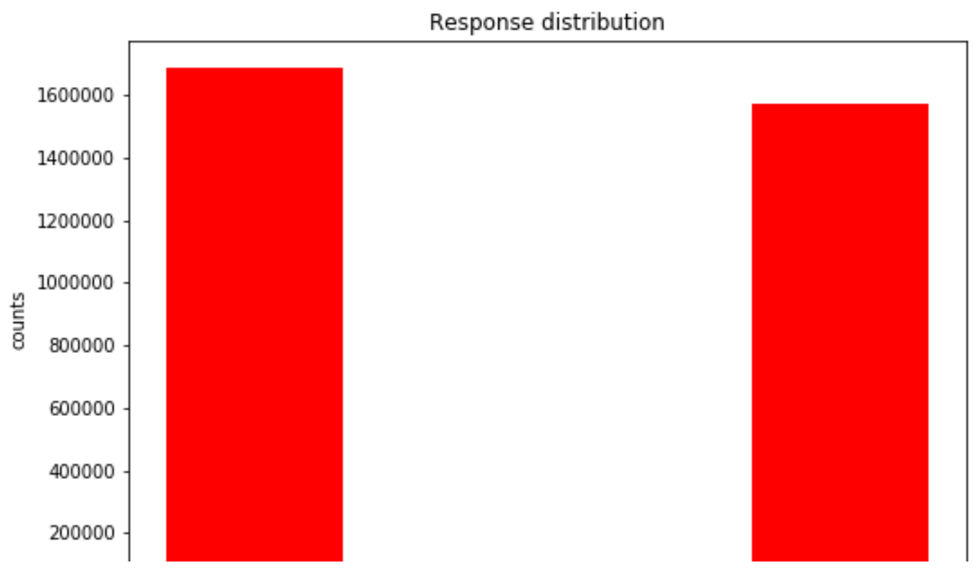
```
#%pyspark
# from matplotlib import pyplot as plt
#import numpy as np
#import functools
# #matplotlib inline

# responses = data.groupBy('HasDetections').count().collect()
# categories = [i[0] for i in responses]
# counts = [i[1] for i in responses]

# ind = np.array(range(len(categories)))
# width = 0.30
# plt.bar(ind, counts, width=width, color='r')

# plt.ylabel('counts')
# plt.title('Response distribution')
# plt.xticks(ind + width/2., categories)
```

```
([<matplotlib.axis.XTick object at 0x7f41b56e5048>, <matplotlib.axis.XTick object at 0x7f41b543ac50>], <a lis
t of 2 Text xticklabel objects>)
```



Printing the count of a target classes.

READY

READY

```
%pyspark
# print(responses)
# print(categories)
# print(counts)
```

```
[Row(HasDetections=1, count=1689470), Row(HasDetections=0, count=1570254)]
[1, 0]
[1689470, 1570254]
```

Droping the Features.

READY

READY

```
%pyspark
data = data.drop('_c0','MachineIdentifier','AvSigVersion')
data.columns
```

```
['ProductName', 'EngineVersion', 'AppVersion', 'IsBeta', 'RtpStateBitfield', 'IsSxsPassiveMode', 'AVProductSt
atesIdentifier', 'AVProductsInstalled', 'AVProductsEnabled', 'HasTpm', 'CountryIdentifier', 'CityIdentifier',
'OrganizationIdentifier', 'GeoNameIdentifier', 'LocaleEnglishNameIdentifier', 'Platform', 'Processor', 'OsVer
', 'OsBuild', 'OsSuite', 'OsPlatformSubRelease', 'OsBuildLab', 'SkuEdition', 'IsProtected', 'AutoSampleOptIn'
, 'SMode', 'IeVerIdentifier', 'SmartScreen', 'Firewall', 'UacLuaenable', 'Census_MDC2FormFactor', 'Census_Dev
iceFamily', 'Census_OEMNameIdentifier', 'Census_OEMModelIdentifier', 'Census_ProcessorCoreCount', 'Census_Pro
cessorManufacturerIdentifier', 'Census_ProcessorModelIdentifier', 'Census_PrimaryDiskTotalCapacity', 'Census_
PrimaryDiskTypeName', 'Census_SystemVolumeTotalCapacity', 'Census_HasOpticalDiskDrive', 'Census_TotalPhysical
RAM', 'Census_ChassisTypeName', 'Census_InternalPrimaryDiagonalDisplaySizeInInches', 'Census_InternalPrimaryD
isplayResolutionHorizontal', 'Census_InternalPrimaryDisplayResolutionVertical', 'Census_PowerPlatformRoleName
', 'Census_InternalBatteryNumberOfCharges', 'Census_OSVersion', 'Census_OSArchitecture', 'Census_OSBranch', '
Census_OSBuildNumber', 'Census_OSBuildRevision', 'Census_OSEdition', 'Census_OSSkuName', 'Census_OSInstallTyp
eName', 'Census_OSInstallLanguageIdentifier', 'Census_OSUILocaleIdentifier', 'Census_OSWUAutoUpdateOptionsNam
e', 'Census_IsPortableOperatingSystem', 'Census_GenuineStateName', 'Census_ActivationChannel', 'Census_IsFlig
htsDisabled', 'Census_FlightRing', 'Census_FirmwareManufacturerIdentifier', 'Census_FirmwareVersionIdentifier
', 'Census_IsSecureBootEnabled', 'Census_IsVirtualDevice', 'Census_IsTouchEnabled', 'Census_IsPenCapable', 'C
ensus_IsAlwaysOnAlwaysConnectedCapable', 'Wdft_IsGamer', 'Wdft_RegionIdentifier', 'HasDetections']
```

Printing the count of a categorical and numerical features.

READY

READY

```
%pyspark
# now let's see how many categorical and numerical features we have:
catcolumns = [item[0] for item in data.dtypes if item[1].startswith('string')]
print(str(len(catcolumns)) + '  categorical features')
numcolumns = [item[0] for item in data.dtypes if item[1].startswith('int') | item[1].startswith('double')][:-
print(str(len(numcolumns)) + '  numerical features')
```

```
26  categorical features
48  numerical features
```

Printing categorical columns.

READY

```
%pyspark
#categorical columns
catcolumns=[item[0] for item in data.dtypes if item[1].startswith('string')]  #will select name of column wit
print("cateogrical columns:", catcolumns)
```

READY

```
cateogrical columns: ['ProductName', 'EngineVersion', 'AppVersion', 'Platform', 'Processor', 'OsVer', 'OsPlat
formSubRelease', 'OsBuildLab', 'SkuEdition', 'SmartScreen', 'Census_MDC2FormFactor', 'Census_DeviceFamily', '
Census_PrimaryDiskTypeName', 'Census_ChassisTypeName', 'Census_PowerPlatformRoleName', 'Census_OSVersion', 'C
ensus_OSArchitecture', 'Census_OSBranch', 'Census_OSEdition', 'Census_OSSkuName', 'Census_OSInstallTypeName',
'Census_OSWUAutoUpdateOptionsName', 'Census_GenuineStateName', 'Census_ActivationChannel', 'Census_FlightRing
']
```

Printing numeric columns.

READY

```
%pyspark
### numerical columns
numcolumns = [item[0] for item in data.dtypes if item[1].startswith('int') | item[1].startswith('double')] #w
    or double data type
print("numerical columns:", numcolumns)
```

READY

```
numerical columns: ['IsBeta', 'RtpStateBitfield', 'IsSxsPassiveMode', 'AVProductStatesIdentifier', 'AVProduct
sInstalled', 'AVProductsEnabled', 'HasTpm', 'CountryIdentifier', 'CityIdentifier', 'OrganizationIdentifier',
'GeoNameIdentifier', 'LocaleEnglishNameIdentifier', 'OsBuild', 'OsSuite', 'IsProtected', 'AutoSampleOptIn', '
SMode', 'IeVerIdentifier', 'Firewall', 'UacLuaenable', 'Census_OEMNameIdentifier', 'Census_OEMModelIdentifier
', 'Census_ProcessorCoreCount', 'Census_ProcessorManufacturerIdentifier', 'Census_ProcessorModelIdentifier',
'Census_PrimaryDiskTotalCapacity', 'Census_SystemVolumeTotalCapacity', 'Census_HasOpticalDiskDrive', 'Census_
TotalPhysicalRAM', 'Census_InternalPrimaryDiagonalDisplaySizeInInches', 'Census_InternalPrimaryDisplayResolut
ionHorizontal', 'Census_InternalPrimaryDisplayResolutionVertical', 'Census_InternalBatteryNumberOfCharges', '
Census_OSBuildNumber', 'Census_OSBuildRevision', 'Census_OSInstallLanguageIdentifier', 'Census_OSUILocaleIden
tifier', 'Census_IsPortableOperatingSystem', 'Census_IsFlightsDisabled', 'Census_FirmwareManufacturerIdentifi
er', 'Census_FirmwareVersionIdentifier', 'Census_IsSecureBootEnabled', 'Census_IsVirtualDevice', 'Census_IsTo
uchEnabled', 'Census_IsPenCapable', 'Census_IsAlwaysOnAlwaysConnectedCapable', 'Wdft_IsGamer', 'Wdft_RegionId
entifier', 'HasDetections']
```

Verifying the type of a numeric columns

READY

```
%pyspark
type(numcolumns)
```

READY

```
<class 'list'>
```

Selecting categorical columns

READY

```
%pyspark
categorical = data.select(catcolumns)
categorical.show()
```

READY

```
+------------+-------------+---------------+---------+---------+--------+-------------------+-------------
------+----------+------------+-------------------+------------------+------------------------+---------
-------------+--------------------------+---------------+--------------------+-------------------+------
------------+------------------+----------------------+------------------------------+-----------------
```

```
-----+-----------------------+----------------+
| ProductName|EngineVersion|      AppVersion| Platform|Processor|   OsVer|OsPlatformSubRelease|         OsBu
ildLab|SkuEdition| SmartScreen|Census_MDC2FormFactor|Census_DeviceFamily|Census_PrimaryDiskTypeName|Census_Ch
assisTypeName|Census_PowerPlatformRoleName|Census_OSVersion|Census_OSArchitecture|     Census_OSBranch|  Cens
us_OSEdition|   Census_OSSkuName|Census_OSInstallTypeName|Census_OSWUAutoUpdateOptionsName|Census_GenuineStat
eName|Census_ActivationChannel|Census_FlightRing|
+------------+-------------+---------------+---------+---------+--------+--------------------+-------------
------+----------+------------+---------------------+-------------------+--------------------------+---------
-------------+----------------------------+----------------+---------------------+--------------------+------
------------+------------------+-----------------------+--------------------------------+------------------
-----+-----------------------+----------------+
|win8defender|  1.1.15100.1| 4.18.1807.18075|windows10|      x64|10.0.0.0|                 rs4|17134.1.amd64f
re....|      Home|RequireAdmin|             Desktop|    Windows.Desktop|                       SSD|
Desktop|                     Desktop|  10.0.17134.165|                amd64|         rs4_release|
Core|              CORE|             UUPUpgrade|                        FullAuto|             IS_GENUINE|
OEM:NONSLP|           Retail|
|win8defender|  1.1.15100.1| 4.18.1807.18075|windows10|      x64|10.0.0.0|                 rs4|17134.1.amd64f
re....|       Pro|RequireAdmin|             Desktop|    Windows.Desktop|                       SSD|
Desktop|                     Desktop|  10.0.17134.165|                amd64|         rs4_release|      Profes
sional|       PROFESSIONAL|             UUPUpgrade|                        FullAuto|             IS_GENUINE|
Retail|           Retail|
|win8defender|  1.1.15100.1| 4.18.1807.18075|windows10|      x64|10.0.0.0|                 rs1|14393.0.amd64f
re....|      Home|RequireAdmin|            Notebook|    Windows.Desktop|                       HDD|
Notebook|                     Mobile|    10.0.14393.0|                amd64|         rs1_release|
Core|              CORE|                Upgrade|                        FullAuto|             IS_GENUINE|
Retail|           Retail|
|win8defender|  1.1.15200.1| 4.18.1807.18075|windows10|      x64|10.0.0.0|                 rs4|17134.1.amd64f
re....|       Pro|RequireAdmin|            Notebook|    Windows.Desktop|                       HDD|
Notebook|                     Mobile|  10.0.17134.254|                amd64|         rs4_release|      Profe
ssional|       PROFESSIONAL|                 Update|                        FullAuto|             IS_GENUINE
|                 Retail|           Retail|
|win8defender|  1.1.15100.1| 4.18.1807.18075|windows10|      x64|10.0.0.0|                 rs4|17134.1.amd64f
re....|      Home|         Off|            Notebook|    Windows.Desktop|                       HDD|
Notebook|                     Mobile|  10.0.17134.191|                amd64|         rs4_release|
Core|              CORE|                  Clean|                        FullAuto|             IS_GENUINE|
Retail|           Retail|
|win8defender|  1.1.15200.1| 4.18.1807.18075|windows10|      x64|10.0.0.0|                 rs4|17134.1.amd64f
re....|       Pro|ExistsNotSet|            Notebook|    Windows.Desktop|                       HDD|
Laptop|                     Mobile|  10.0.17134.228|                amd64|         rs4_release|      Profess
ional|       PROFESSIONAL|             UUPUpgrade|                        FullAuto|             IS_GENUINE|
OEM:DM|           Retail|
|win8defender|  1.1.15100.1| 4.18.1807.18075|windows10|      x64|10.0.0.0|                 rs3|16299.15.amd64
fre...|      Home|RequireAdmin|            Notebook|    Windows.Desktop|                       HDD|
Notebook|                     Mobile|  10.0.16299.309|                amd64|         rs3_release|CoreSingleL
anguage|CORE_SINGLELANGUAGE|             UUPUpgrade|                          Notify|             IS_GENUINE
|                 OEM:DM|           Retail|
|win8defender|  1.1.15200.1| 4.18.1807.18075|windows10|      x64|10.0.0.0|                 rs4|17134.1.amd64f
re....|      Home|RequireAdmin|            Notebook|    Windows.Desktop|                       SSD|
Laptop|                     Mobile|  10.0.17134.228|                amd64|         rs4_release|
Core|              CORE|             UUPUpgrade|                        FullAuto|             IS_GENUINE|
OEM:DM|           Retail|
|win8defender|  1.1.15200.1| 4.18.1806.18062|windows10|      x86|10.0.0.0|                 rs4|17134.1.x86fre
.rs...|       Pro|ExistsNotSet|             Desktop|    Windows.Desktop|                       HDD|
Desktop|                     Desktop|  10.0.17134.137|                  x86|         rs4_release|      Profes
sional|       PROFESSIONAL|             UUPUpgrade|                          Notify|             IS_GENUINE|
Retail|           Retail|
|win8defender|  1.1.15200.1| 4.18.1807.18075|windows10|      x86|10.0.0.0|                 rs3|16299.15.x86fr
e.r...|       Pro|RequireAdmin|             Desktop|    Windows.Desktop|               Unspecified|
Desktop|                     Desktop|  10.0.16299.125|                  x86|         rs3_release|      Profes
sional|       PROFESSIONAL|                 Update|                         UNKNOWN|             IS_GENUINE|
Retail|           Retail|
|win8defender|  1.1.15200.1| 4.18.1807.18075|windows10|      x86|10.0.0.0|                 rs1|14393.576.x86f
re....|       Pro|RequireAdmin|             Notebook|    Windows.Desktop|                       HDD|
Notebook|                      Mobile|  10.0.14393.576|                  x86|          rs1_release|      Profe
ssional|       PROFESSIONAL|                  Update|                              Notify|            IS_GENUINE
|                  Retail|            Retail|
|win8defender|  1.1.15100.1|   4.12.16299.15|windows10|      x64|10.0.0.0|                 rs3|16299.431.amd6
4fr...|      Home|ExistsNotSet|             Desktop|    Windows.Desktop|                       HDD|
Desktop|                     Desktop|  10.0.16299.611|                amd64|rs3_release_svc_e...|
Core|                 CORE|                 Upgrade|                      FullAuto|            IS_GENUINE|
Retail|            Retail|
|win8defender|  1.1.15100.1| 4.18.1807.18075|windows10|      x64|10.0.0.0|                 rs4|17134.1.amd64f
re....|      Home|RequireAdmin|             Notebook|    Windows.Desktop|                       HDD|
Notebook|                      Mobile|  10.0.17134.191|                amd64|          rs4_release|
Core|                 CORE|                 Upgrade|                      FullAuto|            IS_GENUINE|
OEM:DM|            Retail|
|win8defender|  1.1.14800.3|4.14.17639.18041|windows10|      x64|10.0.0.0|                 rs3|16299.431.amd6
4fr...|      Home|RequireAdmin|             Notebook|    Windows.Desktop|                       HDD|
```

Checking the unique values in each categorical columns

READY

```
%pyspark
for i in categorical:
    print(i,(categorical.select(i).distinct().count()))
```

READY

```
Column<b'ProductName'> 2
Column<b'EngineVersion'> 57
Column<b'AppVersion'> 101
Column<b'AvSigVersion'> 6856
Column<b'Platform'> 3
Column<b'Processor'> 3
Column<b'OsVer'> 33
Column<b'OsPlatformSubRelease'> 9
Column<b'OsBuildLab'> 418
Column<b'SkuEdition'> 7
Column<b'SmartScreen'> 16
Column<b'Census_MDC2FormFactor'> 11
Column<b'Census_DeviceFamily'> 3
Column<b'Census_PrimaryDiskTypeName'> 4
Column<b'Census_ChassisTypeName'> 40
Column<b'Census_PowerPlatformRoleName'> 9
Column<b'Census_OSVersion'> 359
```

Importing modules for data preprocessing

READY

```
%pyspark
from pyspark.ml.feature import StringIndexer, VectorAssembler
from pyspark.ml.feature import OneHotEncoder
```

READY

Applying one-hot encoding to the categorical columns

READY

```
%pyspark
stages = []
```

READY

READY

```
%pyspark
#data.select('f1', 'f2').show        #for showing features from the dataframe
```

READY

```
%pyspark
assemblerInputs = [c + "_classVec" for c in catcolumns] + numcolumns
assemblerr = VectorAssembler(inputCols=assemblerInputs, outputCol="features")
#data_cp = assemblerr.transform(data_cp)
stages += [assemblerr]
```

READY

```
%pyspark
from pyspark.ml.feature import MinMaxScaler

scaler = MinMaxScaler(inputCol="features", outputCol="scaledFeatures")
stages += [scaler]
```

READY

if you have target feature as categories then only You have to perform this otherwise escape this

READY

```
%pyspark
# Convert label into label indices using the StringIndexer
#label_stringIdx = StringIndexer(inputCol="HasDetections", outputCol="label")
#stages += [label_stringIdx]
```

```
%pyspark
data.columns
```

READY

```
['ProductName', 'EngineVersion', 'AppVersion', 'IsBeta', 'RtpStateBitfield', 'IsSxsPassiveMode', 'AVProductSt
atesIdentifier', 'AVProductsInstalled', 'AVProductsEnabled', 'HasTpm', 'CountryIdentifier', 'CityIdentifier',
'OrganizationIdentifier', 'GeoNameIdentifier', 'LocaleEnglishNameIdentifier', 'Platform', 'Processor', 'OsVer
', 'OsBuild', 'OsSuite', 'OsPlatformSubRelease', 'OsBuildLab', 'SkuEdition', 'IsProtected', 'AutoSampleOptIn'
, 'SMode', 'IeVerIdentifier', 'SmartScreen', 'Firewall', 'UacLuaenable', 'Census_MDC2FormFactor', 'Census_Dev
iceFamily', 'Census_OEMNameIdentifier', 'Census_OEMModelIdentifier', 'Census_ProcessorCoreCount', 'Census_Pro
cessorManufacturerIdentifier', 'Census_ProcessorModelIdentifier', 'Census_PrimaryDiskTotalCapacity', 'Census_
PrimaryDiskTypeName', 'Census_SystemVolumeTotalCapacity', 'Census_HasOpticalDiskDrive', 'Census_TotalPhysical
RAM', 'Census_ChassisTypeName', 'Census_InternalPrimaryDiagonalDisplaySizeInInches', 'Census_InternalPrimaryD
isplayResolutionHorizontal', 'Census_InternalPrimaryDisplayResolutionVertical', 'Census_PowerPlatformRoleName
', 'Census_InternalBatteryNumberOfCharges', 'Census_OSVersion', 'Census_OSArchitecture', 'Census_OSBranch', '
Census_OSBuildNumber', 'Census_OSBuildRevision', 'Census_OSEdition', 'Census_OSSkuName', 'Census_OSInstallTyp
eName', 'Census_OSInstallLanguageIdentifier', 'Census_OSUILocaleIdentifier', 'Census_OSWUAutoUpdateOptionsNam
e', 'Census_IsPortableOperatingSystem', 'Census_GenuineStateName', 'Census_ActivationChannel', 'Census_IsFlig
htsDisabled', 'Census_FlightRing', 'Census_FirmwareManufacturerIdentifier', 'Census_FirmwareVersionIdentifier
', 'Census_IsSecureBootEnabled', 'Census_IsVirtualDevice', 'Census_IsTouchEnabled', 'Census_IsPenCapable', 'C
ensus_IsAlwaysOnAlwaysConnectedCapable', 'Wdft_IsGamer', 'Wdft_RegionIdentifier', 'HasDetections']
```

Executing Pipeline

READY

```
%pyspark
from pyspark.ml import Pipeline

partialPipeline = Pipeline().setStages(stages)
pipeline_data = partialPipeline.fit(data).transform(data)
```

READY

```
%pyspark
pipeline_data.columns
```

READY

```
['ProductName', 'EngineVersion', 'AppVersion', 'IsBeta', 'RtpStateBitfield', 'IsSxsPassiveMode', 'AVProductSt
atesIdentifier', 'AVProductsInstalled', 'AVProductsEnabled', 'HasTpm', 'CountryIdentifier', 'CityIdentifier',
'OrganizationIdentifier', 'GeoNameIdentifier', 'LocaleEnglishNameIdentifier', 'Platform', 'Processor', 'OsVer
', 'OsBuild', 'OsSuite', 'OsPlatformSubRelease', 'OsBuildLab', 'SkuEdition', 'IsProtected', 'AutoSampleOptIn'
, 'SMode', 'IeVerIdentifier', 'SmartScreen', 'Firewall', 'UacLuaenable', 'Census_MDC2FormFactor', 'Census_Dev
iceFamily', 'Census_OEMNameIdentifier', 'Census_OEMModelIdentifier', 'Census_ProcessorCoreCount', 'Census_Pro
cessorManufacturerIdentifier', 'Census_ProcessorModelIdentifier', 'Census_PrimaryDiskTotalCapacity', 'Census_
PrimaryDiskTypeName', 'Census_SystemVolumeTotalCapacity', 'Census_HasOpticalDiskDrive', 'Census_TotalPhysical
RAM', 'Census_ChassisTypeName', 'Census_InternalPrimaryDiagonalDisplaySizeInInches', 'Census_InternalPrimaryD
isplayResolutionHorizontal', 'Census_InternalPrimaryDisplayResolutionVertical', 'Census_PowerPlatformRoleName
', 'Census_InternalBatteryNumberOfCharges', 'Census_OSVersion', 'Census_OSArchitecture', 'Census_OSBranch', '
Census_OSBuildNumber', 'Census_OSBuildRevision', 'Census_OSEdition', 'Census_OSSkuName', 'Census_OSInstallTyp
eName', 'Census_OSInstallLanguageIdentifier', 'Census_OSUILocaleIdentifier', 'Census_OSWUAutoUpdateOptionsNam
e', 'Census_IsPortableOperatingSystem', 'Census_GenuineStateName', 'Census_ActivationChannel', 'Census_IsFlig
htsDisabled', 'Census_FlightRing', 'Census_FirmwareManufacturerIdentifier', 'Census_FirmwareVersionIdentifier
', 'Census_IsSecureBootEnabled', 'Census_IsVirtualDevice', 'Census_IsTouchEnabled', 'Census_IsPenCapable', 'C
ensus_IsAlwaysOnAlwaysConnectedCapable', 'Wdft_IsGamer', 'Wdft_RegionIdentifier', 'HasDetections', 'ProductNa
me_Index', 'ProductName_classVec', 'EngineVersion_Index', 'EngineVersion_classVec', 'AppVersion_Index', 'AppV
ersion_classVec', 'Platform_Index', 'Platform_classVec', 'Processor_Index', 'Processor_classVec', 'OsVer_Inde
x', 'OsVer_classVec', 'OsPlatformSubRelease_Index', 'OsPlatformSubRelease_classVec', 'OsBuildLab_Index', 'OsB
uildLab_classVec', 'SkuEdition_Index', 'SkuEdition_classVec', 'SmartScreen_Index', 'SmartScreen_classVec', 'C
ensus_MDC2FormFactor_Index', 'Census_MDC2FormFactor_classVec', 'Census_DeviceFamily_Index', 'Census_DeviceFam
ily_classVec', 'Census_PrimaryDiskTypeName_Index', 'Census_PrimaryDiskTypeName_classVec', 'Census_ChassisType
Name_Index', 'Census_ChassisTypeName_classVec', 'Census_PowerPlatformRoleName_Index', 'Census_PowerPlatformRo
leName_classVec', 'Census_OSVersion_Index', 'Census_OSVersion_classVec', 'Census_OSArchitecture_Index', 'Cens
us_OSArchitecture_classVec', 'Census_OSBranch_Index', 'Census_OSBranch_classVec', 'Census_OSEdition_Index', '
Census_OSEdition_classVec', 'Census_OSSkuName_Index', 'Census_OSSkuName_classVec', 'Census_OSInstallTypeName_
Index', 'Census_OSInstallTypeName_classVec', 'Census_OSWUAutoUpdateOptionsName_Index', 'Census_OSWUAutoUpdate
OptionsName_classVec', 'Census_GenuineStateName_Index', 'Census_GenuineStateName_classVec', 'Census_Activatio
nChannel_Index', 'Census_ActivationChannel_classVec', 'Census_FlightRing_Index', 'Census_FlightRing_classVec'
, 'features', 'scaledFeatures']
```

```
%pyspark
pipeline_data = pipeline_data.withColumnRenamed('HasDetections','label')
```

READY

```
%pyspark
pipeline_data.select('scaledFeatures').show(1,False)
```

READY

```
+------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
```

```
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------------+
|scaledFeatures
|
+------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------------+
|[1.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,1.0,0.0,1.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0
.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,1.0,0.
0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0
.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,1.0,0
.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.
0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0
,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,1.0,1.0,0.0,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.2,0.0,0.7581109645202934,0.0,0.2,1.0,0.38461538461538464,0.9143649862762493,0.3333333333333
333,0.2135593220338983,0.1702127659574468,0.8958842322871641,0.9696969696969697,1.0,0.0,0.0,0.345177664974619
27,1.0,5.9604644775390625E-8,0.7987626180397265,0.9195448875098771,0.023622047244094488,0.4444444444444444,0.
44078800089545556,1.4027802677724244E-8,0.009953725448731196,0.0,0.004561527176395592,0.14720452937013445,0.1
5629322268326418,0.12510126142807546,1.0,0.8958744478902358,0.007602635580334516,0.15789473684210525,0.159235
6687898089,0.0,0.0,0.12174721189591078,0.7303339582031522,0.0,0.0,0.0,0.0,0.0,0.0,0.14285714285714285,0.0]|
+------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------
```

```
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
```

Splitting the data into train and test data.

READY

```
%pyspark
(trainingData, testData) = pipeline_data.randomSplit([0.7, 0.3], seed=100)
print(trainingData.count())
print(testData.count())
```

READY

```
2283241
976483
```

```
%pyspark
trainingData.groupBy('label').count().show()
```

READY

```
+-------------+-------+
|HasDetections|  count|
+-------------+-------+
|            1|1183829|
|            0|1099412|
+-------------+-------+
```

```
%pyspark
trainingData.columns
```

READY

```
['ProductName', 'EngineVersion', 'AppVersion', 'IsBeta', 'RtpStateBitfield', 'IsSxsPassiveMode', 'AVProductSt
atesIdentifier', 'AVProductsInstalled', 'AVProductsEnabled', 'HasTpm', 'CountryIdentifier', 'CityIdentifier',
'OrganizationIdentifier', 'GeoNameIdentifier', 'LocaleEnglishNameIdentifier', 'Platform', 'Processor', 'OsVer
', 'OsBuild', 'OsSuite', 'OsPlatformSubRelease', 'OsBuildLab', 'SkuEdition', 'IsProtected', 'AutoSampleOptIn'
, 'SMode', 'IeVerIdentifier', 'SmartScreen', 'Firewall', 'UacLuaenable', 'Census_MDC2FormFactor', 'Census_Dev
iceFamily', 'Census_OEMNameIdentifier', 'Census_OEMModelIdentifier', 'Census_ProcessorCoreCount', 'Census_Pro
cessorManufacturerIdentifier', 'Census_ProcessorModelIdentifier', 'Census_PrimaryDiskTotalCapacity', 'Census_
PrimaryDiskTypeName', 'Census_SystemVolumeTotalCapacity', 'Census_HasOpticalDiskDrive', 'Census_TotalPhysical
RAM', 'Census_ChassisTypeName', 'Census_InternalPrimaryDiagonalDisplaySizeInInches', 'Census_InternalPrimaryD
isplayResolutionHorizontal', 'Census_InternalPrimaryDisplayResolutionVertical', 'Census_PowerPlatformRoleName
', 'Census_InternalBatteryNumberOfCharges', 'Census_OSVersion', 'Census_OSArchitecture', 'Census_OSBranch', '
Census_OSBuildNumber', 'Census_OSBuildRevision', 'Census_OSEdition', 'Census_OSSkuName', 'Census_OSInstallTyp
eName', 'Census_OSInstallLanguageIdentifier', 'Census_OSUILocaleIdentifier', 'Census_OSWUAutoUpdateOptionsNam
e', 'Census_IsPortableOperatingSystem', 'Census_GenuineStateName', 'Census_ActivationChannel', 'Census_IsFlig
htsDisabled', 'Census_FlightRing', 'Census_FirmwareManufacturerIdentifier', 'Census_FirmwareVersionIdentifier
', 'Census_IsSecureBootEnabled', 'Census_IsVirtualDevice', 'Census_IsTouchEnabled', 'Census_IsPenCapable', 'C
ensus_IsAlwaysOnAlwaysConnectedCapable', 'Wdft_IsGamer', 'Wdft_RegionIdentifier', 'label', 'ProductName_Index
', 'ProductName_classVec', 'EngineVersion_Index', 'EngineVersion_classVec', 'AppVersion_Index', 'AppVersion_c
lassVec', 'Platform_Index', 'Platform_classVec', 'Processor_Index', 'Processor_classVec', 'OsVer_Index', 'OsV
er_classVec', 'OsPlatformSubRelease_Index', 'OsPlatformSubRelease_classVec', 'OsBuildLab_Index', 'OsBuildLab_
classVec', 'SkuEdition_Index', 'SkuEdition_classVec', 'SmartScreen_Index', 'SmartScreen_classVec', 'Census_MD
C2FormFactor_Index', 'Census_MDC2FormFactor_classVec', 'Census_DeviceFamily_Index', 'Census_DeviceFamily_clas
sVec', 'Census_PrimaryDiskTypeName_Index', 'Census_PrimaryDiskTypeName_classVec', 'Census_ChassisTypeName_Ind
ex', 'Census_ChassisTypeName_classVec', 'Census_PowerPlatformRoleName_Index', 'Census_PowerPlatformRoleName_c
lassVec', 'Census_OSVersion_Index', 'Census_OSVersion_classVec', 'Census_OSArchitecture_Index', 'Census_OSArc
hitecture_classVec', 'Census_OSBranch_Index', 'Census_OSBranch_classVec', 'Census_OSEdition_Index', 'Census_O
SEdition_classVec', 'Census_OSSkuName_Index', 'Census_OSSkuName_classVec', 'Census_OSInstallTypeName_Index',
'Census_OSInstallTypeName_classVec', 'Census_OSWUAutoUpdateOptionsName_Index', 'Census_OSWUAutoUpdateOptionsN
ame_classVec', 'Census_GenuineStateName_Index', 'Census_GenuineStateName_classVec', 'Census_ActivationChannel
_Index', 'Census_ActivationChannel_classVec', 'Census_FlightRing_Index', 'Census_FlightRing_classVec', 'featu
res', 'scaledFeatures']
```

```
%pyspark
testData.groupBy('label').count().show()
```

READY

```
+-------------+------+
|HasDetections| count|
+-------------+------+
|            1|505709|
|            0|471735|
+-------------+------+
```

```
%pyspark
pipeline_data.select('features').show(5)
```

READY

```
+--------------------+
|            features|
+--------------------+
|(8051,[1,3,4,5,6,...|
|(8051,[1,3,4,5,6,...|
|(8051,[1,3,4,5,6,...|
|(8051,[1,3,4,5,6,...|
|(8051,[1,3,4,5,6,...|
+--------------------+
only showing top 5 rows
```

```
%pyspark
print(trainingData.count(),len(trainingData.dtypes))
print(testData.count(),len(testData.dtypes))
#print(trainingData.columns,testData.columns)
```

READY

```
2282280 128
977444 128
```

Model building starts here.

READY

Applying Logistic Regression.

READY

READY

```
%pyspark
from pyspark.ml.classification import LogisticRegression

# Create initial LogisticRegression model
lr = LogisticRegression(labelCol="label", featuresCol="scaledFeatures", maxIter=10)

# Train model with Training Data
Model = lr.fit(trainingData)

train_prediction = Model.transform(trainingData)

test_prediction = Model.transform(testData)
```

Evaluating the model performance.

READY

READY

```
%pyspark
from pyspark.ml.evaluation import MulticlassClassificationEvaluator

evaluator = MulticlassClassificationEvaluator(labelCol="label", predictionCol="prediction", metricName="accur
train_accuracy = evaluator.evaluate(train_prediction)
test_accuracy = evaluator.evaluate(test_prediction)
print("Accuracy of LogisticRegression on train is = %g"% (train_accuracy))
print("Accuracy of LogisticRegression on test is = %g"% (test_accuracy))
#print("Test Error of LogisticRegression = %g " % (1.0 - lr_accuracy_test))
```

```
Accuracy of LogisticRegression on train is = 0.9261
Accuracy of LogisticRegression on test is = 0.925656
```

Confusion Matrix.

READY

READY

```
%pyspark
cm  = test_prediction.crosstab('prediction','label')
cm = cm.toPandas()
cm
```

```
  prediction_label       0       1
0              1.0   42802  476109
1              0.0  427778   29794
```

READY

```
%pyspark

TP = cm["1"][0]
FP = cm["0"][0]
TN = cm["0"][1]
FN = cm["1"][1]
print(TP,FP,TN,FN)
```

```
476109 42802 427778 29794
```

Calculating LR model Accuracy, Sensitivity, Specificity, Precision.

READY

READY

```
%pyspark

Accuracy = (TP+TN)/(TP+TN+FP+FN)
Sensitivity = TP/(TP+FN)
Specificity = TN/(TN+FP)
Precision = TP/(TP+FP)
```

```
ACCURACY = 0.93
SENSITIVITY = 0.94
SPECIFICITY = 0.91
PRECISION = 0.92
```

Receiver operating curve

READY

READY

```
%pyspark
from pyspark.ml.evaluation import BinaryClassificationEvaluator

#predictions_LR = lrModel.transform(test)
evaluator = BinaryClassificationEvaluator()
print("Test_SET (Area Under ROC): " + str(evaluator.evaluate(test_prediction, {evaluator.metricName: "areaUnc
```

```
Test_SET (Area Under ROC): 0.9755296969167018
```

Applying Decision Tree.

READY

READY

```
%pyspark
from pyspark.ml.classification import DecisionTreeClassifier

dt = DecisionTreeClassifier(labelCol="label", featuresCol="scaledFeatures",maxDepth=3)
Model = dt.fit(trainingData)
train_prediction = Model.transform(trainingData)
test_prediction = Model.transform(testData)
```

Evaluating the Decision Tree.

READY

READY

```
%pyspark
evaluator = MulticlassClassificationEvaluator(labelCol="label", predictionCol="prediction", metricName="acc
train_accuracy = evaluator.evaluate(train_prediction)
test_accuracy = evaluator.evaluate(test_prediction)
print("Accuracy of DecisionTree on train is = %g"% (train_accuracy))
print("Accuracy of DecisionTree on test is = %g"% (test_accuracy))
#print("Test Error of DecisionTree = %g " % (1.0 - dt_accuracy_test))
```

```
Accuracy of DecisionTree on train is = 1
Accuracy of DecisionTree on test is = 1
```

READY

```
%pyspark
cm  = test_prediction.crosstab('prediction','label')
cm = cm.toPandas()
cm
```

```
  prediction_label       0       1
0              1.0       0  505903
1              0.0  470580       0
```

Calculating DT model Accuracy, Sensitivity, Specificity, Precision.

READY

READY

```
%pyspark

TP = cm["1"][0]
FP = cm["0"][0]
TN = cm["0"][1]
FN = cm["1"][1]
print(TP,FP,TN,FN)

Accuracy = (TP+TN)/(TP+TN+FP+FN)
Sensitivity = TP/(TP+FN)
Specificity = TN/(TN+FP)
Precision = TP/(TP+FP)
print('ACCURACY = %0.2f' %Accuracy)
print('SENSITIVITY = %0.2f' %Sensitivity)
print('SPECIFICITY = %0.2f' %Specificity)
print('PRECISION = %0.2f' %Precision)
```

```
505903 0 470580 0
ACCURACY = 1.00
SENSITIVITY = 1.00
SPECIFICITY = 1.00
PRECISION = 1.00
```

READY

```
%pyspark
from pyspark.ml.evaluation import BinaryClassificationEvaluator

evaluator = BinaryClassificationEvaluator()
print("Test_SET (Area Under ROC): " + str(evaluator.evaluate(test_prediction, {evaluator.metricName: "areaUnc
```

```
Test_SET (Area Under ROC): 1.0
```

Applying Naive Bayes.

READY

READY

```
%pyspark
from pyspark.ml.classification import NaiveBayes

nb = NaiveBayes(labelCol="label", featuresCol="scaledFeatures")
Model = nb.fit(trainingData)
train_prediction = Model.transform(trainingData)
test_prediction = Model.transform(testData)
```

Evaluating the NB model.

READY

READY

```
%pyspark
evaluator = MulticlassClassificationEvaluator(labelCol="label", predictionCol="prediction", metricName="acc
train_accuracy = evaluator.evaluate(train_prediction)
test_accuracy = evaluator.evaluate(test_prediction)
print("Accuracy of DecisionTree on train is = %g"% (train_accuracy))
print("Accuracy of DecisionTree on test is = %g"% (test_accuracy))
#print("Test Error of DecisionTree = %g " % (1.0 - dt_accuracy_test))
```

```
Accuracy of DecisionTree on train is = 0.913496
Accuracy of DecisionTree on test is = 0.913599
```

Calculating NB model Accuracy, Sensitivity, Specificity, Precision.

READY

```
%pyspark

cm  = test_prediction.crosstab('prediction','label')
cm = cm.toPandas()
cm
```

READY

```
  prediction_label       0       1
0             1.0   84369  505903
1             0.0  386211       0
```

```
%pyspark

TP = cm["1"][0]
FP = cm["0"][0]
TN = cm["0"][1]
FN = cm["1"][1]
print(TP,FP,TN,FN)

Accuracy = (TP+TN)/(TP+TN+FP+FN)
Sensitivity = TP/(TP+FN)
Specificity = TN/(TN+FP)
Precision = TP/(TP+FP)
print('ACCURACY = %0.2f' %Accuracy)
print('SENSITIVITY = %0.2f' %Sensitivity)
print('SPECIFICITY = %0.2f' %Specificity)
print('PRECISION = %0.2f' %Precision)
```

READY

```
%pyspark
from pyspark.ml.evaluation import BinaryClassificationEvaluator

evaluator = BinaryClassificationEvaluator()
print("Test_SET (Area Under ROC): " + str(evaluator.evaluate(test_prediction, {evaluator.metricName: "areaUnc
```

READY

Applying Random Forest.

READY

```
%pyspark
from pyspark.ml.classification import RandomForestClassifier

Rf = RandomForestClassifier(labelCol = "label",featuresCol = "scaledFeatures")
Model = Rf.fit(trainingData)
train_prediction = Model.transform(trainingData)
test_prediction = Model.transform(testData)
```

READY

Calculating RF model Accuracy, Sensitivity, Specificity, Precision.

READY

```
%pyspark
evaluator = MulticlassClassificationEvaluator(labelCol="label", predictionCol="prediction", metricName="acc
train_accuracy = evaluator.evaluate(train_prediction)
test_accuracy = evaluator.evaluate(test_prediction)
print("Accuracy of DecisionTree on train is = %g"% (train_accuracy))
print("Accuracy of DecisionTree on test is = %g"% (test_accuracy))
#print("Test Error of DecisionTree = %g " % (1.0 - dt_accuracy_test))
```

READY

```
Accuracy of DecisionTree on train is = 0.7514
Accuracy of DecisionTree on test is = 0.751561
```

```
%pyspark

cm  = test_prediction.crosstab('prediction','label')
cm = cm.toPandas()
cm
```

READY

```
%pyspark

TP = cm["1"][0]
FP = cm["0"][0]
TN = cm["0"][1]
FN = cm["1"][1]
print(TP,FP,TN,FN)

Accuracy = (TP+TN)/(TP+TN+FP+FN)
Sensitivity = TP/(TP+FN)
Specificity = TN/(TN+FP)
Precision = TP/(TP+FP)
print('ACCURACY = %0.2f' %Accuracy)
print('SENSITIVITY = %0.2f' %Sensitivity)
print('SPECIFICITY = %0.2f' %Specificity)
print('PRECISION = %0.2f' %Precision)
```

READY

```
%pyspark
from pyspark.ml.evaluation import BinaryClassificationEvaluator

evaluator = BinaryClassificationEvaluator()
print("Test_SET (Area Under ROC): " + str(evaluator.evaluate(test_prediction, {evaluator.metricName: "areaUnc
```

READY