In [1]:

```
#Implementation of Random Forest using Sklearn and CuML
```

In [2]:

```
import cudf
import numpy as np
import pandas as pd

from cuml.ensemble import RandomForestClassifier as curfc
from cuml.metrics import accuracy_score

from sklearn.ensemble import RandomForestClassifier as skrfc
from sklearn.datasets import make_classification
from sklearn.model_selection import train_test_split
```

In [3]:

```
#Define Parameters
n_samples = 2**18
n_features = 399
n_info = 300
data_type = np.float32
```

In [4]:

```
%%time
#Generate Data
X,y = make_classification(n_samples=n_samples,
                          n_features=n_features,
                          n_informative=n_info,
                          random_state=123, n_classes=2)

X = pd.DataFrame(X.astype(data_type))
y = pd.Series(y)

X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size = 0.2,
                                                    random_state=0)
```

```
CPU times: user 33.9 s, sys: 820 ms, total: 34.7 s
Wall time: 34.7 s
```

In [5]:

```
%%time
#Convert to Cudf
X_cudf_train = cudf.DataFrame.from_pandas(X_train)
X_cudf_test = cudf.DataFrame.from_pandas(X_test)
y_cudf_train = cudf.Series(y_train.values)
y_cudf_test = cudf.Series(y_test.values)
```

```
CPU times: user 1.98 s, sys: 752 ms, total: 2.73 s
Wall time: 2.74 s
```

In [6]:

```
%%time
#SCikitlearn Model
sk_model = skrfc(n_estimators=35,
                 max_depth=15,
                 max_features=1.0,
                 random_state=23)
```

```
sk_model.fit(X_train, y_train)
```

```
CPU times: user 47min 40s, sys: 660 ms, total: 47min 40s
Wall time: 47min 41s
```

Out[6]:

```
RandomForestClassifier(max_depth=15, max_features=1.0, n_estimators=35,
                       random_state=23)
```

In [7]:

```
%%time
#Evaluate
sk_predict = sk_model.predict(X_test)
sk_acc = accuracy_score(y_test, sk_predict)
print('accuracy is',sk_acc)
```

```
accuracy is 0.8743062019348145
CPU times: user 556 ms, sys: 4 ms, total: 560 ms
Wall time: 559 ms
```

In [8]:

```
%%time
#CUML Model
cuml_model = curfc(n_estimators=35,
                   max_depth=15,
                   max_features=1.0,
                   random_state=23)

cuml_model.fit(X_cudf_train, y_cudf_train)
```

```
/opt/conda/envs/rapids/lib/python3.7/site-packages/cuml/internals/api_decorators.py:794:
UserWarning: For reproducible results in Random Forest Classifier or for almost reproduc
ible results in Random Forest Regressor, n_streams==1 is recommended. If n_streams is >
1, results may vary due to stream/thread timing differences, even when random_state is s
et
  return func(**kwargs)
CPU times: user 5min 29s, sys: 689 ms, total: 5min 30s
Wall time: 52.8 s
```

Out[8]:

```
RandomForestClassifier()
```

In [9]:

```
%%time
#Evaluate
Pred_y = cuml_model.predict(X_cudf_test)

cuml_accuracy = accuracy_score(y_cudf_test, Pred_y)
print('accuracy is ', cuml_accuracy)
```

```
accuracy is  0.8727231025695801
CPU times: user 10.1 s, sys: 192 ms, total: 10.3 s
Wall time: 168 ms
```

In [ ]: