Challenge 8

In [1]:
```
import pandas as pd
import numpy as np
from sklearn import preprocessing
from sklearn.svm import LinearSVC, NuSVC
from sklearn.linear_model import SGDClassifier
from sklearn.neighbors import KNeighborsClassifier, NearestCentroid
from sklearn.model_selection import cross_val_score
import warnings
warnings.filterwarnings('ignore')
```

Import Labeled and Unlabeled Datasets:

In [2]:
```
labeled = pd.read_csv('Dataset_Challenge8/8_labeled.csv')
x_unlabeled = pd.read_csv('Dataset_Challenge8/8_unlabeled.csv')
```

Split Labeled dataset into X_train and y_train:

In [3]:
```
x_train = labeled.iloc[:,:-1]
y_train = labeled.iloc[:,-1]
```

Encoding categorical columns to numrical:

In [4]:
```
le = preprocessing.LabelEncoder()
columns = x_train.columns.tolist()
for x in columns:
    x_train[x] = le.fit_transform(x_train[x])
    x_unlabeled[x] = le.fit_transform(x_unlabeled[x])
```

Define a list of Models:

In [5]:
```
models = [
    LinearSVC(),
    NuSVC(nu=0.0001),
    SGDClassifier(max_iter=1000, tol=1e-3),
    NearestCentroid(),
    KNeighborsClassifier()
]
```

Train the models on labeled data with 5 fold cross validation and RMSE scoring:

In [6]:
```
for model in models:
    model.seed = 42
    num_folds = 5
    scores = cross_val_score(model, x_train, y_train, cv=num_folds, scoring='neg_mean_s
    score_description = " %0.2f (+/- %0.2f)" % (np.sqrt(scores.mean()*-1), scores.std()
    print('{model:25} CV-5 RMSE: {score}'.format(model=model.__class__.__name__, score=
```

```
LinearSVC                 CV-5 RMSE:  0.62 (+/- 0.09)
NuSVC                     CV-5 RMSE:  0.74 (+/- 0.20)
SGDClassifier             CV-5 RMSE:  0.69 (+/- 0.31)
NearestCentroid           CV-5 RMSE:  0.57 (+/- 0.04)
KNeighborsClassifier      CV-5 RMSE:  0.49 (+/- 0.03)
```

Use the trained models to predict labels for unlabaled data and retrain the

models on merged dataset:

In [7]:
```
for model in models:
    model.seed = 42
    num_folds = 5
    model.fit(x_train, y_train)   # Training
    y_unlabeled = model.predict(x_unlabeled)  # Creating pseudo-labeled data
    y_unlabeled = pd.DataFrame(y_unlabeled, columns = ['malware'])
    seudo_labeled_data = x_unlabeled.join(y_unlabeled)
    seudo_labeled_data = seudo_labeled_data.fillna(0)
    x_merged = x_train.append(seudo_labeled_data.iloc[:,:-1])
    y_merged = y_train.append(seudo_labeled_data.iloc[:,-1])
    scores = cross_val_score(model, x_merged.sample(frac=1,random_state=10), y_merged.s
                             cv=num_folds, scoring='neg_mean_squared_error')
    score_description = " %0.2f (+/- %0.2f)" % (np.sqrt(scores.mean()*-1), scores.std()
    print('{model:25} CV-5 RMSE: {score}'.format(model=model.__class__.__name__, score=
```

```
LinearSVC                 CV-5 RMSE:  0.49 (+/- 0.57)
NuSVC                     CV-5 RMSE:  0.47 (+/- 0.76)
SGDClassifier             CV-5 RMSE:  0.24 (+/- 0.05)
NearestCentroid           CV-5 RMSE:  0.45 (+/- 0.01)
KNeighborsClassifier      CV-5 RMSE:  0.14 (+/- 0.00)
```