In [41]:
```
#CSDA1050 Advanced Analytics Capstone Course

# Project Sprint 3
# Improving student's graduation in Education
# By Sylvain Kamto
# Student ID: 11060
```

In [1]:
```
import pandas as pd
import collections
import random
import scipy as sp
import numpy as np
import collections
from sklearn.decomposition import FactorAnalysis
from sklearn.ensemble import RandomForestRegressor, GradientBoostingC
lassifier, RandomForestClassifier
from sklearn.model_selection import train_test_split, cross_val_score
from matplotlib import pyplot as plt
import re
from matplotlib import pyplot as plt
plt.style.use('ggplot')
import copy
pd.options.mode.chained_assignment = None  # default='warn'
```

In [28]:
```
#Import Data
```

In [2]:
```
assessments = pd.read_csv('assessments.csv')
courses = pd.read_csv('courses.csv')
studentass = pd.read_csv('studentAssessment.csv')
studentinfo = pd.read_csv('studentinfo.csv')
studentreg = pd.read_csv('studentRegistration.csv')
studentvle = pd.read_csv('studentVle.csv')
vle = pd.read_csv('vle.csv')
```

In [29]:
```
# Data comes in 7 different files. I merge data to do preliminary ana
lyses
```

In [3]:
```
dset = pd.merge(assessments, studentass, how = 'inner', on = 'id_asse
ssment')
dset =pd.merge(dset, studentinfo, how = 'inner', on = ['code_module',
'code_presentation', 'id_student'])
dset = pd.merge(dset, courses, how ='inner', on =['code_module', 'cod
e_presentation'])
dset = pd.merge(dset, studentreg, how = 'inner', on = ['code_module',
'code_presentation', 'id_student'])
ass_nofinal = assessments.loc[assessments.assessment_type != 'Exam']
```