Module 4: Introduction to NumPy & Pandas

Case Study Solution

edureka!

Case Study Solution

Domain – Education

focus – Data analysis

Business challenge/requirement
You are a data analyst with University of Cal USA (Not a machine learning expert yet as you still have not completed ML with Python Course :-)). The University has data of Math, Physics and Data Structure score of sophomore students. This data is stored in different files. The University has hired a data science company to do analysis of scores and find if there is any correlation of score with age, ethnicity etc. Before the data is given to the company you have to do data wrangling.

Key issues
Ensure students identify is not revealed to the agency and only relevant data is shared

Considerations
NONE

Data volume
- In thousands, but only around 1800 records are shared in files MathScoreTerm1.csv DSScoreTerm1.csv, PhysicsScoreTerm1.csv

Additional information
- NA

Business benefits
University can get more students enrollment by improving its international ranking through personalized course/curriculum for students

Approach to Solve
You have to use fundamentals of Numpy and Pandas covered in module 4.

1. Read the three csv files which contains the score of same students in term1 for each Subject

2. Remove the name and ethnicity column (to ensure confidentiality)
3. Fill missing score data with zero
4. Merge the three files
5. Change Sex(M/F) Column to 1/2 for further analysis
6. Store the data in new file – ScoreFinal.csv

Enhancements for code

You can try these enhancements in code

1. Convert ethnicity to numerical value
2. Fill the missing score for a student to the average of the class

Solution

```
import pandas as pd
import numpy as np


# Read the three files

math_data = pd.read_csv("MathScoreTerm1.csv")
ds_data = pd.read_csv("DSScoreTerm1.csv")
physics_data = pd.read_csv("PhysicsScoreTerm1.csv")

# Preview the first data set
print(math_data.head())


#Remove Name and Ethinicty columns from each  data set
math_data = math_data.drop(['Name','Ethinicity'], axis=1)
print(math_data.head())

# print Summary of missing values in Term1 Data
print (math_data.isnull().sum())

# Fill Missing Scores with 0
math_data['Score'] = math_data['Score'].fillna(0)
# Check Again if values are filled correctly
print (math_data.isnull().sum())
```

```
# Do the same for DS and Physics

ds_data = ds_data.drop(['Name','Ethinicity'], axis=1)
ds_data['Score'] = ds_data['Score'].fillna(0)

physics_data = physics_data.drop(['Name','Ethinicity'], axis=1)
physics_data['Score'] = physics_data['Score'].fillna(0)

all_data =[math_data,ds_data,physics_data]

# Convert Sex M to 1 and F to 2
for dataset in all_data:
    dataset['Sex'] = dataset['Sex'].map({'M': 1, 'F': 2}).astype(int)

all_data_df = pd.concat(all_data)
all_data_df.to_csv('ScoreFinal.csv',index=False)

print ("Done")
```