MACHINE LEARNING

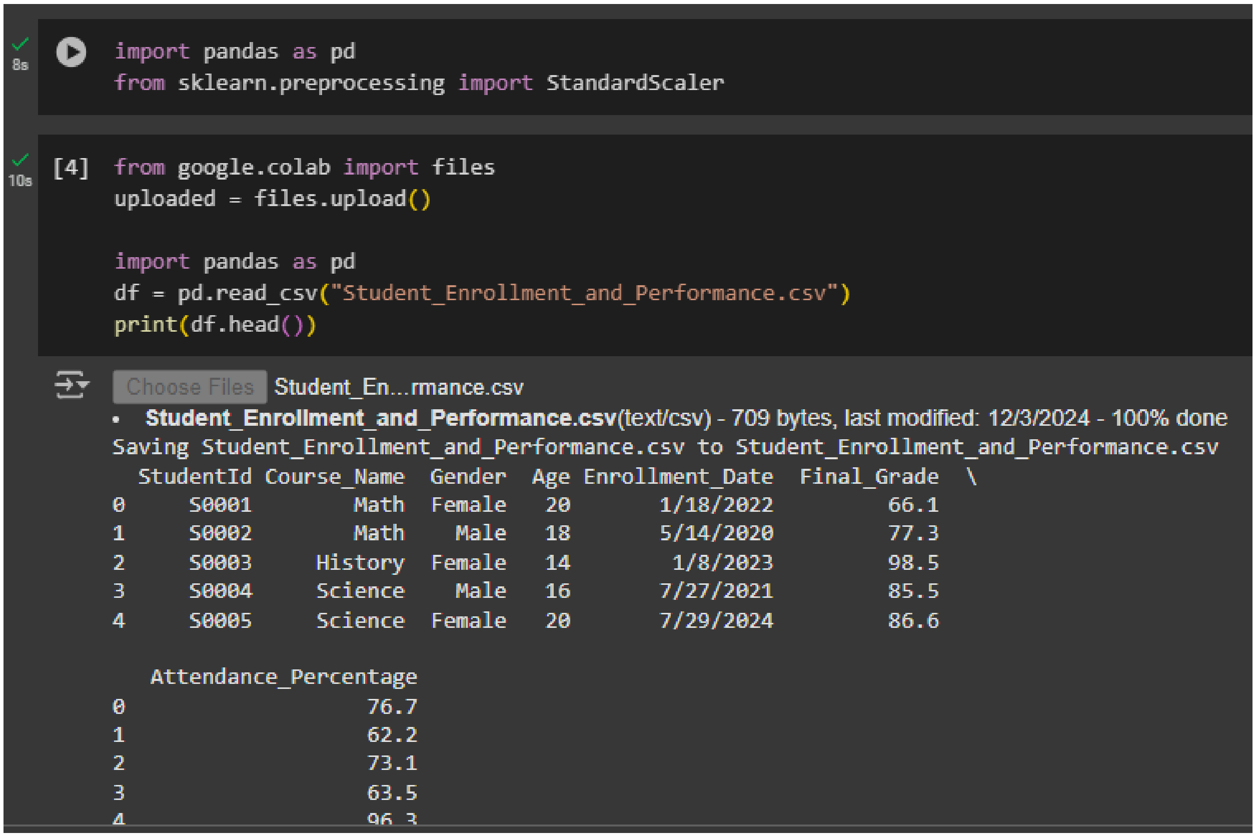
LAB 3

NIKKI EVANA BLESSY.N

2341459

Aggregation, Scaling and data wrangling

On the datasets of your choice apply the concepts of data wrangling, data combination and data scaling.



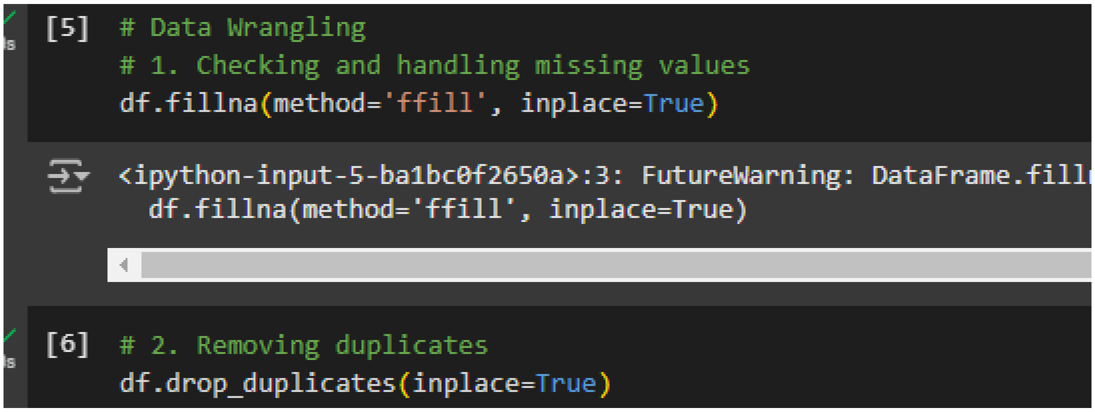
```
import pandas as pd
from sklearn.preprocessing import StandardScaler
```

```
[4] from google.colab import files
    uploaded = files.upload()

    import pandas as pd
    df = pd.read_csv("Student_Enrollment_and_Performance.csv")
    print(df.head())
```

Choose Files Student_En...rmance.csv

- **Student_Enrollment_and_Performance.csv**(text/csv) - 709 bytes, last modified: 12/3/2024 - 100% done

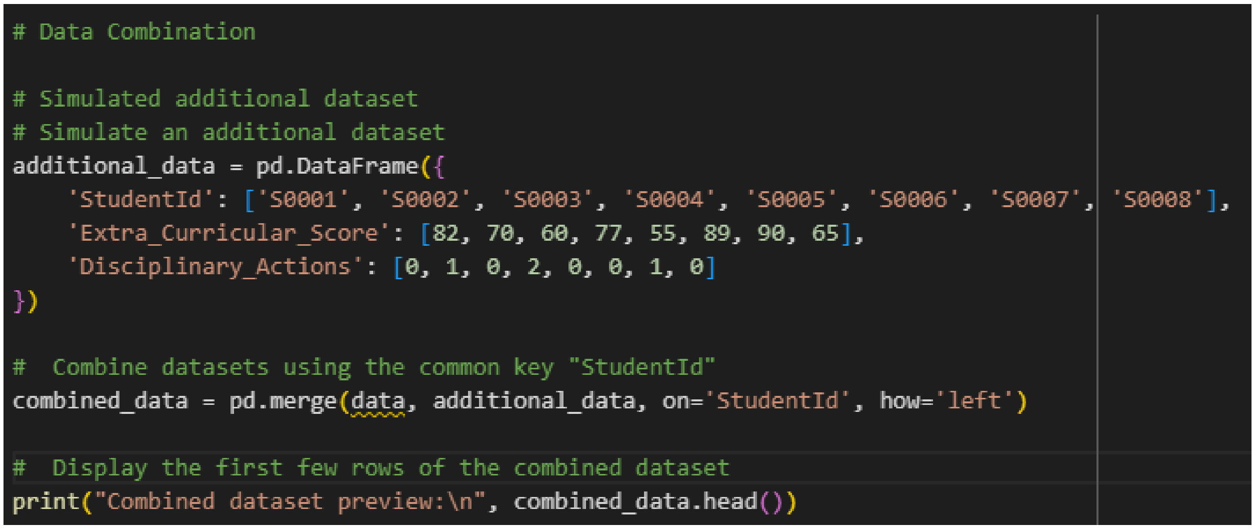
```
Saving Student_Enrollment_and_Performance.csv to Student_Enrollment_and_Performance.csv
  StudentId Course_Name  Gender  Age Enrollment_Date  Final_Grade  \
0     S0001        Math  Female   20       1/18/2022         66.1
1     S0002        Math    Male   18       5/14/2020         77.3
2     S0003     History  Female   14        1/8/2023         98.5
3     S0004     Science    Male   16       7/27/2021         85.5
4     S0005     Science  Female   20       7/29/2024         86.6

   Attendance_Percentage
0                   76.7
1                   62.2
2                   73.1
3                   63.5
4                   96.3
```



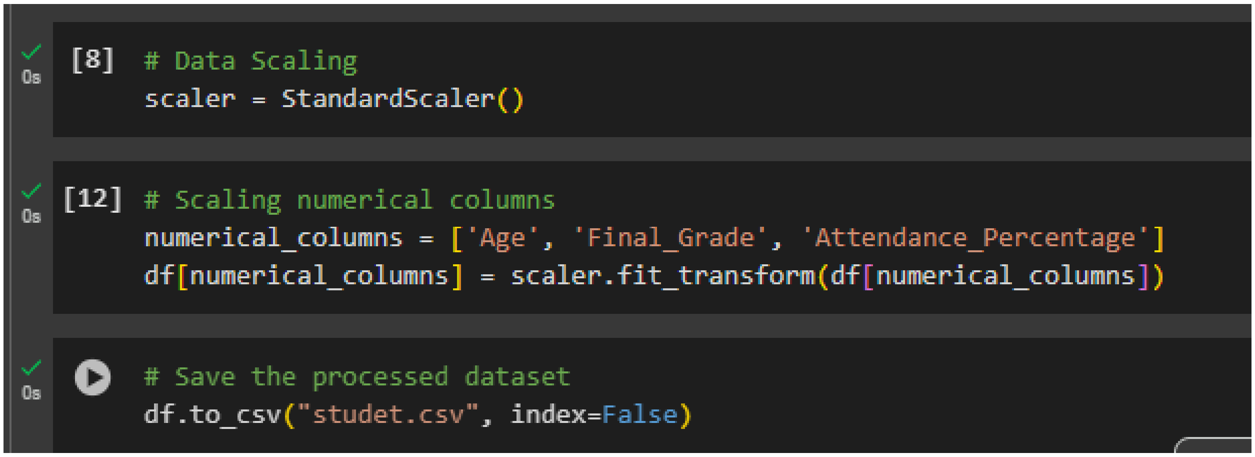
```
# Data Wrangling
# 1. Checking and handling missing values
df.fillna(method='ffill', inplace=True)
```

```
<ipython-input-5-ba1bc0f2650a>:3: FutureWarning: DataFrame.fillı
  df.fillna(method='ffill', inplace=True)
```

```
# 2. Removing duplicates
df.drop_duplicates(inplace=True)
```



```
# Data Combination

# Simulated additional dataset
# Simulate an additional dataset
additional_data = pd.DataFrame({
    'StudentId': ['S0001', 'S0002', 'S0003', 'S0004', 'S0005', 'S0006', 'S0007', 'S0008'],
    'Extra_Curricular_Score': [82, 70, 60, 77, 55, 89, 90, 65],
    'Disciplinary_Actions': [0, 1, 0, 2, 0, 0, 1, 0]
})

#  Combine datasets using the common key "StudentId"
combined_data = pd.merge(data, additional_data, on='StudentId', how='left')

#  Display the first few rows of the combined dataset
print("Combined dataset preview:\n", combined_data.head())
```



```
# Data Scaling
scaler = StandardScaler()
```

```
# Scaling numerical columns
numerical_columns = ['Age', 'Final_Grade', 'Attendance_Percentage']
df[numerical_columns] = scaler.fit_transform(df[numerical_columns])
```

```
# Save the processed dataset
df.to_csv("studet.csv", index=False)
```