

Cleaning of The Dataset (Student Classification Dataset)

1.1 مقدمة

تحتوي مجموعة البيانات الخاصة بالطلاب على البيانات التي تمثل عوامل مؤثرة في معدلاتهم ومستوياتهم الدراسية، لكن تلك المجموعة لم تكن نقية وجاهزة لبدء تحليلها والعمل عليها، من هذا المنطلق اضطررت الى تنقيتها.

1.2 الأدوات المستخدمة

- اللغة البرمجية بايثون.
- مكتبة بانداز.
- محرر اكواد جوبيتر او فيجيوال ستوديو كود.

1.3 الذي تم تنظيفه

تم حذف جميع الصفوف التي توجد فيها قيم فارغة، أيضاً تم حذف الاسطر التي يوجد فيها قيم غير صحيحة وتم شرحها في هذا الجدول:

جدول (1.1) توضيح ما تم تنظيفه

اسم العمود	رقم الصف	القيمة التي حُذف بسببها الصف	القيم الافتراضية للعمود
Student_Age	14	37	From 18 to 25
Transport	84	Car	Private, Bus
Sex	137	96	Female, Male
Notes	60	6	Yes, No
Attendance	112	3	Always, Never, or Sometimes
Listening_in_Class	126	6	Yes or No

1.4 الخطوات

سلسلة الخطوات البرمجية المتبعة حتى اكتمال التنظيف والحصول على النتيجة النهائية:

- 1- استدعاء المكتبة، وتحويل الملف الخاص بمجموعة البيانات الى داتا فريم، ثم طباعة المتغير.

```
Click here to ask Blackbox to help you code faster |
import pandas as pd
df = pd.read_csv('student.csv')
print(df.to_string())
✓ 0.0s
```

صورة (1.1) الخطوة الأولى

- 2- حذف جميع الصفوف التي تحتوي على قيم فارغة وطباعة الملف المحدث.

```
Click here to ask Blackbox to help you code faster |
new_df = df.dropna()
print(new_df.to_string())
✓ 0.0s
```

صورة (1.2) الخطوة الثانية

- 3- التأكد من كل القيم الخاصة بعمود عمر الطالب؛ حيث لا تقل عن 18 ولا تزيد عن 25، وان وجد ما دون ذلك يتم حذف الصف، (تم حذف القيمة 37 في هذه الخطوة).

```
Click here to ask Blackbox to help you code faster |
for x in new_df.index:
    if ~(18 <= new_df.loc[x, 'Student_Age'] <= 25):
        new_df.drop(x, inplace=True)
print(new_df.to_string())
✓ 0.0s
```

صورة (1.3) الخطوة الثالثة

4- التأكد من جميع القيم الخاصة بعمود جنس الطالب، وان لا تخرج القيم عن ذكر او انثى، واي بيانات أخرى يتم حذف صفها، (تم حذف القيمة "96" في هذه الخطوة).

```
Click here to ask Blackbox to help you code faster |
for x in new_df.index:
    if 'Male'!= new_df.loc[x, 'Sex'] != 'Female':
        new_df.drop(x, inplace=True)
print(new_df.to_string())
✓ 0.0s
```

صورة (1.4) الخطوة الرابعة

5- التأكد من جميع القيم الخاصة بعمود طريقة تنقل الطالب، وان لا تخرج القيم عن خاص او باص، واي بيانات أخرى يتم حذف صفها، (تم حذف القيمة "Car" في هذه الخطوة).

```
Click here to ask Blackbox to help you code faster |
for x in new_df.index:
    if 'Bus'!= new_df.loc[x, 'Transportation'] != 'Private':
        new_df.drop(x, inplace=True)
print(new_df.to_string())
✓ 0.0s
```

صورة (1.5) الخطوة الخامسة

6- التأكد من جميع القيم الخاصة بعمود هل يكتب الطالب ملاحظات، وان لا تخرج القيم عن نعم او لا، واي بيانات أخرى يتم حذف صفها، (تم حذف القيمة "6" في هذه الخطوة).

```
Click here to ask Blackbox to help you code faster |
for x in new_df.index:
    if 'No'!= new_df.loc[x, 'Notes'] != 'Yes':
        new_df.drop(x, inplace=True)
print(new_df.to_string())
✓ 0.0s
```

صورة (1.6) الخطوة السادسة

7- التأكد من جميع القيم الخاصة بعمود حضور الطالب، وان لا تخرج القيم عن دائما او احيائاً او ابداء، واي بيانات أخرى يتم حذف صفها، (تم حذف القيمة "3" في هذه الخطوة).

```
Click here to ask Blackbox to help you code faster |
for x in new_df.index:
    if new_df.loc[x, 'Attendance'] != 'Sometimes' and new_df.loc[x, 'Attendance'] != 'Never' and new_df.loc[x, 'Attendance'] != 'Always':
        new_df.drop(x, inplace=True)
print(new_df.to_string())
✓ 0.0s
```

صورة (1.7) الخطوة السابعة

8- التأكد من جميع القيم الخاصة بعمود هل يركز الطالب اثناء الدروس، وان لا تخرج القيم عن نعم او لا، واي بيانات أخرى يتم حذف صفها، (تم حذف القيمة "6" في هذه الخطوة).

```
Click here to ask Blackbox to help you code faster |
for x in new_df.index:
    if 'No' != new_df.loc[x, 'Listening_in_Class'] != 'Yes':
        new_df.drop(x, inplace=True)
print(new_df.to_string())
✓ 0.0s
```

صورة (1.8) الخطوة الثامنة

9- تحويل الداتا فريم المنقى الى ملف سي اس في.

```
Click here to ask Blackbox to help you code faster |
new_df.to_csv('CleanedFile.csv')
✓ 0.0s
```

صورة (1.9) الخطوة التاسعة