

درس: علم داده

عنوان: پروژه نهایی (علم داده در تحصیل)

نام: علی پورقیصری

استاد: محمد‌مهدی رضاپور

شماره دانشجویی: 993613014

فهرست مطالب

[1- مقدمه 2](#_Toc160795605)

[2- مجموعه داده 3](#_Toc160795606)

[3- مدل سازی 3](#_Toc160795607)

[1-3- الگوریتم K-means 3](#_Toc160795608)

[2-3- الگوریتم K-medians 5](#_Toc160795609)

[3-3- الگوریتم DBSCAN 5](#_Toc160795610)

[مراجع 7](#_Toc160795611)

1- مقدمه

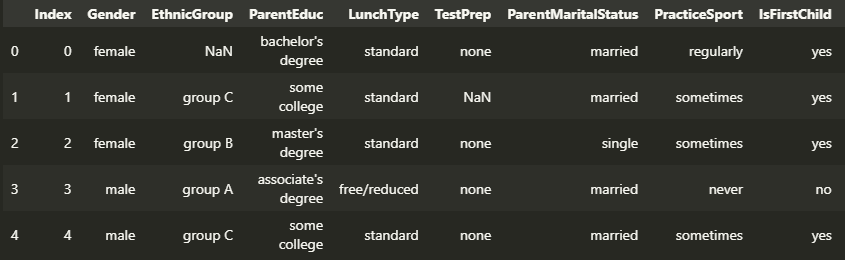
خوشه‌ بندی یا Clustering تکنیکی است که شامل گروه‌بندی اشیاء مشابه بر اساس شباهت‌های ذاتی آن‌ها می‌شود. به عبارت دیگر، هدف آن این است که نقاط داده را به خوشه‌های مجزا تقسیم کند، جایی که نقاط درون یک خوشه بیشتر به یکدیگر شباهت دارند تا به خوشه‌های دیگر. با کشف این گروه‌بندی‌های طبیعی، الگوریتم‌های خوشه‌بندی می‌توانند بینش‌های ارزشمندی را در مورد ساختار زیربنایی داده‌ها ارائه دهند.

2- مجموعه داده

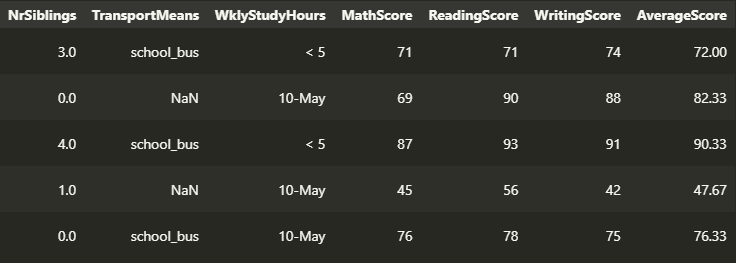
این مجموعه داده به نام Students test score: Extended Dataset یک مجموعه داده با داده های بیشتر برای مجموعه داده اولیه می‌باشد. این مجموعه داده شامل 30642 سطر و 16 ستون می‌باشد که به صورت زیر است.

1. شماره سطر
2. جنسیت
3. قومیت
4. تحصیلات والدین
5. نوع مدرسه
6. آزمون آمادگی
7. وضعیت تاهل والدین
8. تمرین ورزشی
9. آیا اولین فرزند است؟
10. تعداد خواهران / برادران
11. وسیله حمل و نقل
12. مجموع ساعات درس خواندن هفتگی
13. نمره ریاضی
14. نمره خواندن
15. نمره نوشتن
16. نمره میانگین

این مجموعه داده نیاز به تمیز سازی و پیش پردازش دارد زیرا داده های خالی در بین این داده ها وجود دارد. این نوع داده ها را با چند روش می‌توان از بین برد که در ادامه توضیحات کافی را برای این بخش می‌دهیم.

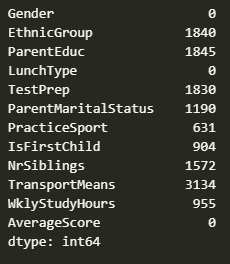
در اینجا می‌توانیم محتویات 5 سطر اول را ببینیم تا به یک درکی از داده برسیم.

1- پنج سطر اول مجموعه داده



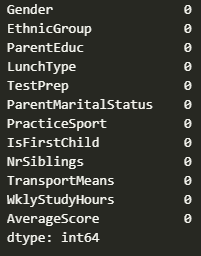
2- پنج سطر اول مجموعه داده

حال ستون‌هایی که به درد ما نمی‌خورند را حذف می‌کنیم. این ستون‌ها شامل شماره سطر، نمرات ریاضی، خواندن و نوشتن است زیرا برچسب نهایی که میانگین است، با احتساب این سه سطر به وجود آمده است.

بعد از حذف این چهار ستون به سراغ داده‌های از دست رفته می‌رویم. اگر داده‌های از دست رفته را نشان دهیم، به صورت زیر هستند.

3- مجموعه داده های از دست رفته

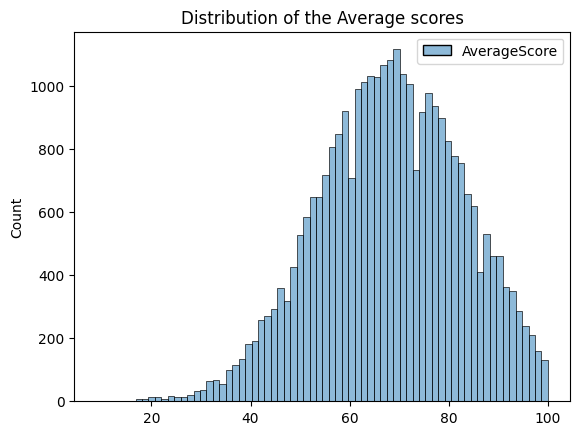
این داده‌ها را در اینجا با استفاده از پر تکرار ترین مقدار، پر می‌کنیم. پس از این کار می‌توان دید که دیگر داده گم شده نداریم.



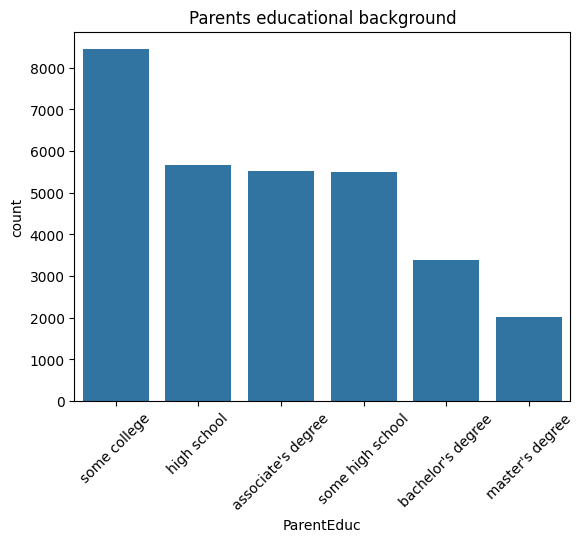
4- پر کردن داده های از دست رفته

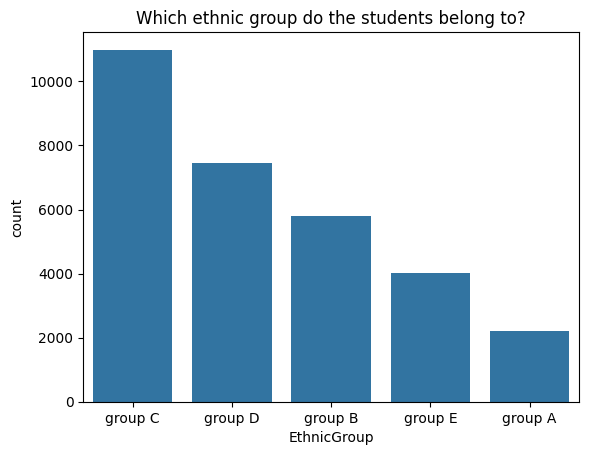
حال تعداد سطر های تکراری را اندازه گیری می‌کنیم و می‌بینیم که 132 سطر تکراری داریم. در اینجا اقدام به حذف این داده‌ها می‌کنیم.

حال می‌توانیم نمودار هایی برای ستون های مختلف رسم کنیم تا درک بیشتری پیدا کنیم.



5- توزیع میانگین نمرات



6- تحصیلات والدین

7- قومیت

مراجع

* [Understanding DBSCAN Algorithm and Implementation from Scratch](https://towardsdatascience.com/understanding-dbscan-algorithm-and-implementation-from-scratch-c256289479c5)
* [Clustering with DBSCAN, Clearly Explained!!! - YouTube](https://www.youtube.com/watch?v=RDZUdRSDOok)
* [K-Means نحوه کار الگوریتم](https://raahbord.com/k-means/)