دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مدیریت علم و فناوری

گزارش کار هفته نهم

داده های پرت

نگارش

رضا اکبری مقدم

استاد

دکتر مهدی قطعی

دی ماه ۹۹

فهرست مطالب

مقدمه ۴

مجموعه داده ۴

گزارش کار ۵

نتیجه گیری..... ۱۳

منابع ۱۳

فهرست اشکال

شکل ۱ ..... ۴
شکل ۲ ..... ۵
شکل ۳ ..... ۶
شکل ۴ ..... ۷
شکل ۵ ..... ۷
شکل ۶ ..... ۸
شکل ۷ ..... ۸
شکل ۸ ..... ۹
شکل ۹ ..... ۹
شکل ۱۰ ..... ۹
شکل ۱۱ ..... ۱۰
شکل ۱۲ ..... ۱۰
شکل ۱۳ ..... ۱۱
شکل ۱۴ ..... ۱۱
شکل ۱۵ ..... ۱۲
شکل ۱۶ ..... ۱۲
شکل ۱۷ ..... ۱۲

مقدمه

داده پرت یا داده دورافتاده(Outlier) در مبحث آمار، به داده‌ای گفته می‌شود که با دیگر داده‌های هم‌گروه فاصله چشمگیری داشته‌باشد. داده‌های پَرت همه‌جا هستند. آن‌ها به دلایل مختلفی تولید می‌شوند و معمولاً در میان انواع داده‌ها دیده می‌شوند. این نوع داده‌ها را که معمولاً غیرعادی هستند و از الگوهای عمومی در یک مجموعه‌ی داده پیروی نمی‌کنند، می‌توان توسط الگوریتم‌های مختلفِ تشخیصِ داده‌های پرت شناسایی کرد. با شناسایی داده‌های پَرت می‌توان آن‌ها را از مجموعه‌ی داده کنار گذاشت تا مجموعه‌ی داده، کمی تمیزتر و مناسب‌تر جهت تزریق به الگوریتم‌هایی مانند طبقه‌بندی و خوشه‌بندی باشد. البته در برخی از مواقع خودِ داده‌های پَرت هستند که صورت مسئله می‌باشند. مثلاً در بین بیماران و علائمِ آن‌ها ممکن است به دنبال بیمارانی بگردیم که علائمشان با دیگر بیماران همخوانی ندارد و به نوعی در آن مجموعه‌ی داده، غیر طبیعی هستند.

مجموعه داده

مجموعه داده satellite شامل ۶۴۳۵ نمونه و ۳۶ ویژگی میباشد که این ویژگی ها بیانگر ۴ نوع طیف رنگی میباشد که یک ماتریس ۳*۳ را تشکیل میدهد. همچنین این مجموعه داده یک ستون جداگانه outlier دارد که یک تشخیص از داده های پرت میباشد و بایستی بررسی شود تا چه میزان این تشخیص صحیح میباشد.

	0	1	2	3	4	5	6	7	8	9	...	26	27	28	29	30	31	32	33	34	35
0	92	115	120	94	84	102	106	79	84	102	...	134	104	88	121	128	100	84	107	113	87
1	84	102	106	79	84	102	102	83	80	102	...	128	100	84	107	113	87	84	99	104	79
2	84	102	102	83	80	102	102	79	84	94	...	113	87	84	99	104	79	84	99	104	79
3	80	102	102	79	84	94	102	79	80	94	...	104	79	84	99	104	79	84	103	104	79
4	84	94	102	79	80	94	98	76	80	102	...	104	79	84	103	104	79	79	107	109	87
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6430	60	83	96	85	64	87	100	88	64	83	...	104	92	66	87	108	89	63	83	104	85
6431	64	79	100	85	56	71	96	85	56	68	...	100	85	66	83	100	85	63	83	100	81
6432	56	68	91	81	56	64	91	81	53	64	...	100	81	59	87	96	81	63	83	92	74
6433	56	68	87	74	60	71	91	81	60	64	...	96	74	59	83	92	74	59	83	92	70
6434	60	71	91	81	60	64	104	99	56	64	...	92	74	59	83	92	70	63	79	108	92

6435 rows × 36 columns

شکل ۱

گزارش کار

در این گزارش از ابزار پایتون جهت تشخیص داده های پرت استفاده شده است.

پیش پردازش

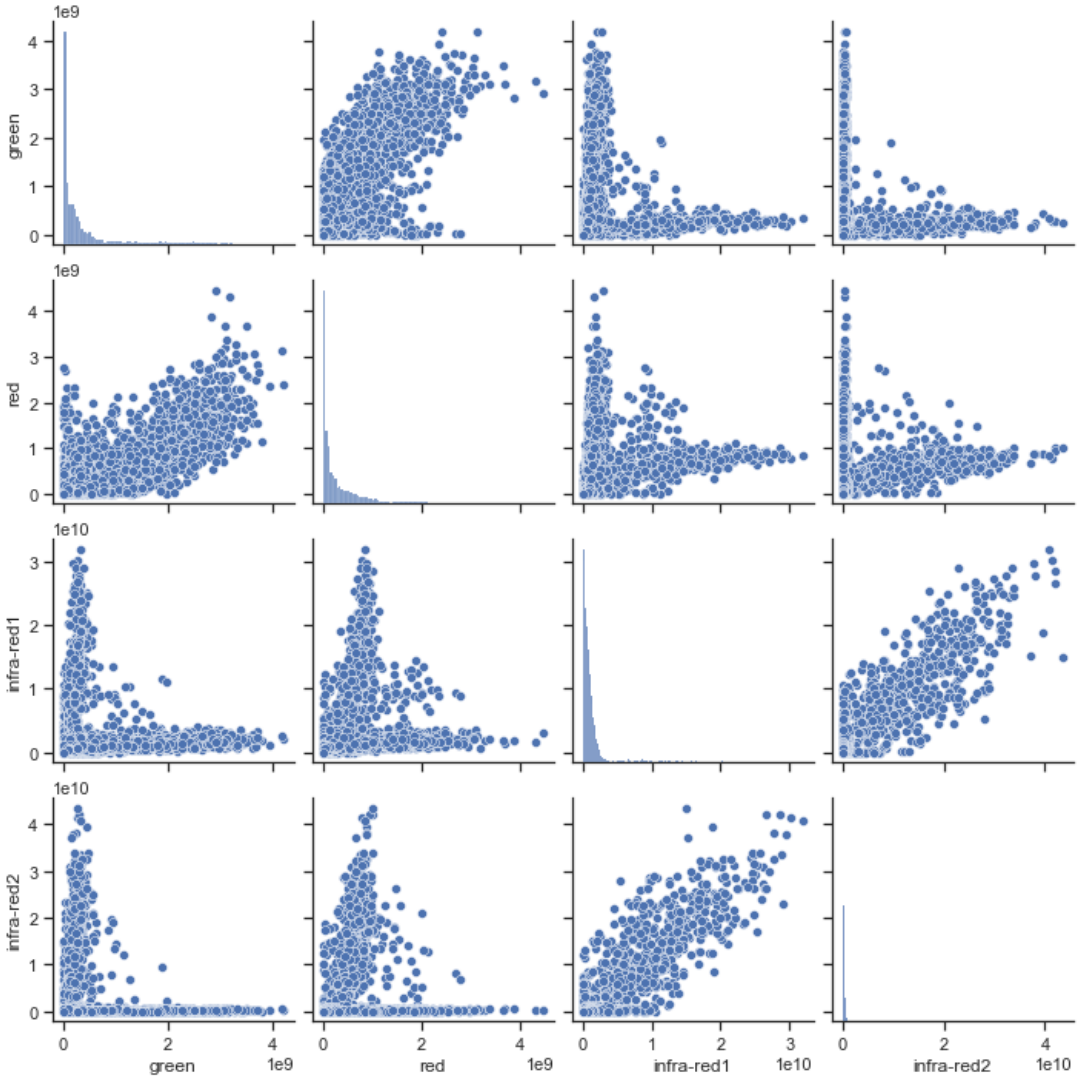
به جهت اینکه داده ها را کوچک کنیم و ابعاد جدول را کاهش دهیم با استفاده از فرمول دترمینان ماتریس مقدار دترمینان هر ماتریس را بدست میاوریم تا ابعاد جدول به ۴ بعد کاهش پیدا کند و همچنین نام ۴ طیف رنگی را بر روی این بعد ها قرار میدهیم. شکل ۲ جدول خروجی این فرآیند را نمایش میدهد.

	green	red	infra-red1	infra-red2
0	92368368	358976352	1641722200	65235456
1	256075344	269012352	1560819104	347076900
2	150380496	200644974	989791400	541577175
3	39874880	106617150	813613736	491776975
4	46497024	367058118	825290856	238630876
...	...	...	...	...
6430	760416000	70404003	840420100	94407548
6431	537868864	42500	705554016	29089125
6432	274996736	4460544	686940000	3175524
6433	259100856	25725184	743775312	4551296
6434	-178560	-31168	224540	-26936

6435 rows × 4 columns

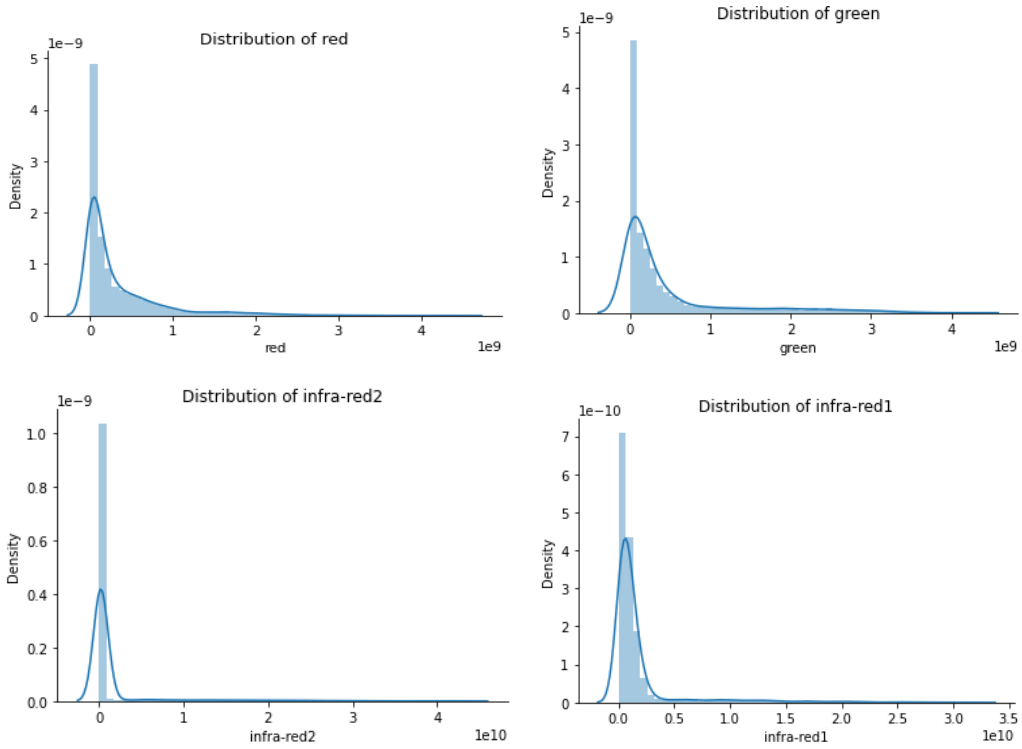
شکل ۲

سپس با استفاده از نمودارهای پراکندگی و جعبه ای داده ها را مصورسازی میکنیم.



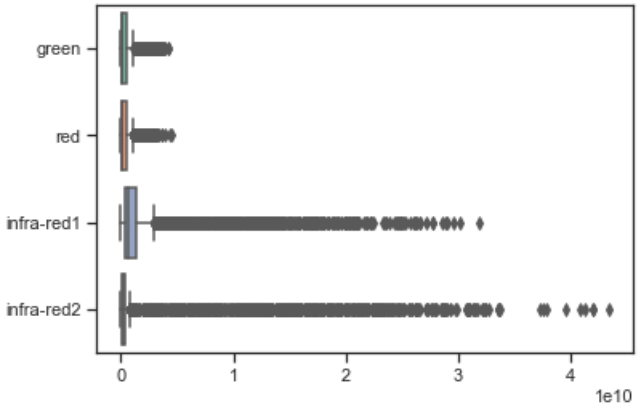
شکل ۳

در شکل ۳ مشاهده میشود ماتریس نمودار پراکندگی داده ها نمایش داده شده است.



شکل ۴

در شکل ۴ نمودار پراکندگی هر ویژگی به صورت جداگانه نمایش داده شده است. در این نمودار میتوان مشاهده نمود که اکثر داده ها در محدوده ۰ تا ۰/۵ تجمع دارند.



شکل ۵

در شکل ۵ نمودار جعبه ای ۴ ویژگی این مجموعه داده نمایش داده شده است که نشان میدهد اکثر تجمع داده ها در قسمت صفر تا یک میباشد.

تشخیص و حذف داده های پرت

ابتدا ضریب چولگی و ضریب کشیدگی ویژگی red را قبل از تشخیص داده های پرت محاسبه میکنیم.

```
count    6.435000e+03
mean     3.441616e+08
std      5.178838e+08
min     -3.116800e+04
25%      2.778027e+07
50%      1.190696e+08
75%      4.487905e+08
max      4.460618e+09
Name: red, dtype: float64
```

شکل ۶

در شکل ۶ توصیفی از این ویژگی از جمله میانه و مینیمم و ماکزیمم و چارک ها و... نمایش داده شده است.

در تئوری احتمال و آمار، چولگی بیانگر میزان عدم تقارن توزیع احتمال داده‌ها حول میانگینشان است. مقدار چولگی می‌تواند منفی یا مثبت باشد .

همچنین ضریب کشیدگی جهت نشان دادن عدم همخوانی قله یا نوک منحنی برای بعضی از توزیع‌های آماری نسبت به توزیع نرمال استفاده میشود.

```
Skewness: 2.494742
Kurtosis: 7.334154
```

شکل ۷

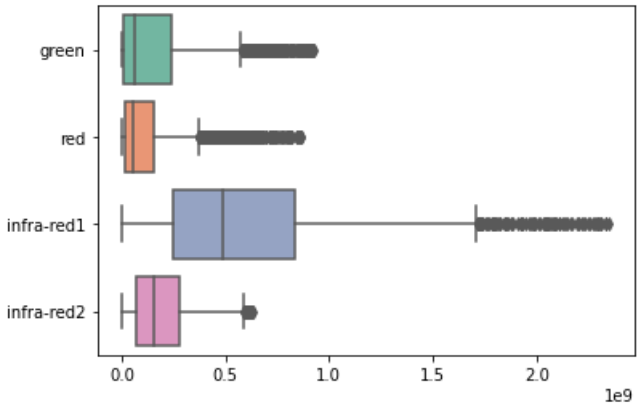
در شکل ۷ مشاهده میشود که ضریب چولگی ۲,۴۹۴ بیانگر این میباشد ویژگی red ما به سمت چپ نمودار پراکندگی جمع شده است و ضریب کشیدگی ۷,۳۳۴ میباشد که نشان میدهد که نوک منحنی کشیده میباشد. که در مجموع میتوان نتیجه گرفت مجموعه داده دارای داده پرت میباشد.

اولین روش محاسبه داده های پرت با استفاده از نمودار جعبه ای و چارک های روی مجموعه داده نقطه میانی بین چارک اول و سوم را برای تمام ویژگی ها محاسبه میکنیم.

```
green          4.511514e+08
red            4.210103e+08
infra-red1     1.019517e+09
infra-red2     2.770072e+08
```

شکل ۸

در شکل ۸ خروجی این محاسبات را مشاهده میکنید. سپس این مقادیر را به چارک سوم اضافه و از چارک اول کم میکنیم و داده های خارج این بازه را به عنوان داده پرت حساب کرده و حذف میکنیم.



شکل ۹

همانطور که در شکل ۹ مشاهده میشود فضای محدوده مجموعه داده از ۵ به ۲ کاهش پیدا کرد. همچنین تعداد نمونه های ما از ۶۴۳۵ به ۴۴۷۴ نمونه کاهش پیدا کرد.

```
Skewness: 2.046781
Kurtosis: 4.403925
```

شکل ۱۰

در شکل ۱۰ ضریب چولگی و کشیدگی مجموعه داده پس از حذف داده های پرت نشان داده شده است که میتوان مشاهده کرد به میزان قابل توجهی ضریب کشیدگی کاهش پیدا کرده است.

در ادامه با استفاده از ضریب z score به تشخیص داده های پرت میپردازیم.

مقدار Z-score از طریق رابطه زیر محاسبه می‌شود که در آن، μمقدار میانگین جمعیت آماری و σ انحراف معیار جمعیت می‌باشد. مقدار قدر مطلق (absolute value) محاسبه شده برایz ، فاصله آن ردیف از داده ها

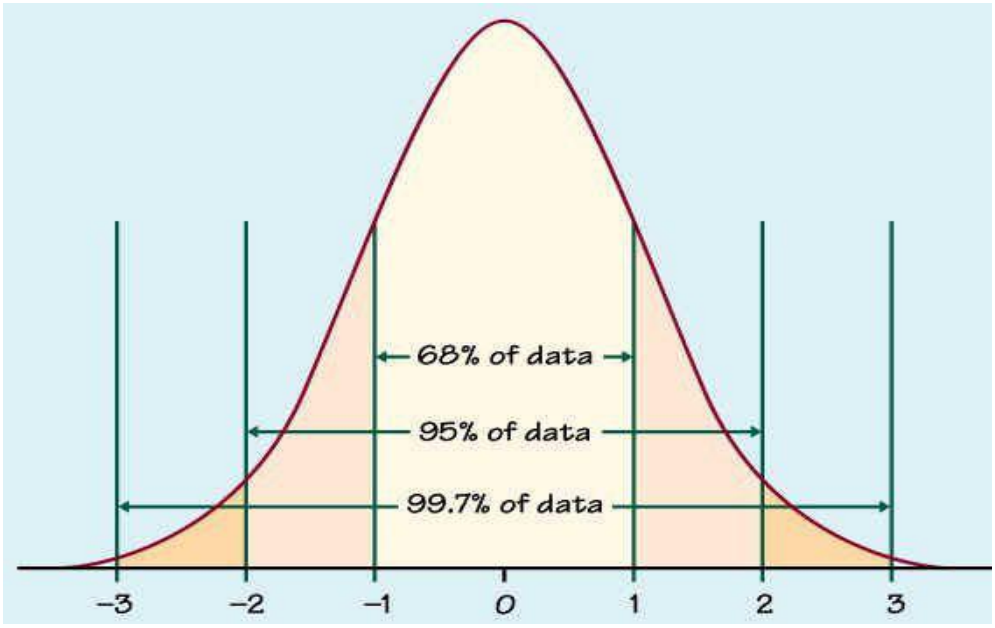
را از میانگین کل جمعیت بر حسب انحراف معیار نشان می‌دهد. هنگامی که این مقدار مثبت باشد، یعنی Z-score بالاتر از میانگین و اگر منفی باشد، نشان دهنده کمتر بود آن مقدار خاص، از میانگین کل داده ها می‌باشد.

$$z = \frac{x - \mu}{\sigma}$$

$$\mu = \text{Mean}$$
$$\sigma = \text{Standard Deviation}$$

شکل ۱۱

براساس ویژگی های توزیع نرمال مطابق شکل زیر، ۹۹,۷٪ داده ها در فاصله‌ی ۶ انحراف معیار از میانگین قرار، دارند. لذا می‌توان نتیجه گرفت که ۰,۳٪ داده ها که خارج از این حدود قرار می‌گیرند، رفتاری نامتعارف نسبت به اکثریت داده ها دارند.



شکل ۱۲

بر این اساس با استفاده از این مقدار به بررسی داده های پرت میپردازیم که داده هایی که مقدار z score آن ها بین ۳ و ۳- میباشد را انتخاب میکنیم.

	green	red	infra-red1	infra-red2	zscore
0	92368368	358976352	1641722200	65235456	0.028606
1	256075344	269012352	1560819104	347076900	-0.145108
2	150380496	200644974	989791400	541577175	-0.277121
3	39874880	106617150	813613736	491776975	-0.458683
4	46497024	367058118	825290856	238630876	0.044212
...	...	...	...	...	...
6430	760416000	70404003	840420100	94407548	-0.528608
6431	537868864	42500	705554016	29089125	-0.664472
6432	274996736	4460544	686940000	3175524	-0.655941
6433	259100856	25725184	743775312	4551296	-0.614880
6434	-178560	-31168	224540	-26936	-0.664614

6251 rows × 5 columns

شکل ۱۳

در شکل ۱۳ میتوان مشاهده نمود که نمونه های ما از ۶۴۳۵ به ۶۲۵۱ نمونه کاهش پیدا کرده است.

```
Skewness: 1.939797
Kurtosis: 3.519661
```

شکل ۱۴

در شکل ۱۴ ضریب کشیدگی و چولگی ویژگی red محاسبه شده است که نشان میدهد به چه میزان قابل توجهی این معیار تاثیر مثبتی در کاهش کشیدگی و چولگی داده های ما شده است با توجه به اینکه نمونه های خیلی کمتری از این مجموعه داده حذف شده است.

در ادامه با استفاده از الگوریتم IsolationForest به تشخیص داده های پرت میپردازیم.

ایده اصلی در الگوریتم جنگل ایزوله، که متفاوت از سایر روش‌های تشخیص ناهنجاری‌ها است، در نحوه برخورد با داده‌های آموزشی است. در این روش به جای بررسی و ایجاد پروفایل نقاط عادی و متعارف، صریحاً مشاهدات ناهنجار مشخص می‌شوند. جنگل ایزوله، مانند هر روش ساختار درختی (Tree Structure) دیگر، بر اساس درخت تصمیم‌گیری (Decision Tree) ساخته شده است. در این درختان، ابتدا تقسیم‌بندی‌هایی با انتخاب

تصادفی یک ویژگی و سپس تعیین مقدار تقسیم تصادفی بین حداقل و حداکثر مقدار آن ویژگی انتخاب شده، صورت می‌گیرد.

پس از اعمال این الگوریتم بر روی مجموعه داده ابتدا به بررسی مقدار accuracy این الگوریتم میپردازیم.

```
IsolationForest Accuracy: 0.7752913752913753
```

شکل ۱۵

در شکل ۱۵ مقدار دقت این الگوریتم حدودا ۷۷ درصد میباشد که مقدار قابل قبولی است.

یکی دیگر از الگوریتم های محاسبه و تشخیص داده های پرت LocalOutlierFactor میباشد.

این الگورتیم برای کشف ناهنجاری‌ها و داده‌های پرت موجود در نقاط داده با اندازه‌گیری انحراف محلی یک نقطه داده با توجه به همسایه‌های آن ارائه شد.

همانند روش های قبلی داده ها را به مجموعه تست و آموزش تقسیم میکنیم و سپس الگوریتم را بر روی آن ها اعمال میکنیم تا ببینیم پیشبینی الگوریتم به چه صورتی میباشد.

```
LocalOutlierFactor Accuracy: 0.9813519813519813
```

شکل ۱۶

همانطور که در شکل ۱۶ مشاهده میشود میزان دقت این الگوریتم ۹۸ درصد میباشد که تقریبا اکثر داده های پرت را شناسایی و تشخیص داده است.

الگوریتم بعدی که مقایسه میکنیم الگوریتم میباشد.

الگوریتک Elliptical Envelope در یک داده توزیع شده Gaussian ، نقاط دورافتاده را تشخیص می دهد.

```
EllipticEnvelope Accuracy: 0.98989898989899
```

شکل ۱۷

در شکل ۱۷ مشاهده میشود که دقت این الگوریتم تقریبا ۹۹ درصد میباشد که بالاترین دقت را در بین الگوریتم های بررسی شده دارا میباشد.

نتیجه گیری

با بررسی های انجام شده در این مجموعه داده به این نتیجه رسیدیم که الگوریتم ها و روش های تشخیص و حذف داده های پرت همیشه به طور قطع به بهبود عملکرد مجموعه داده کمک نمیکنند و ممکن است در مواردی به بدتر شدن وضعیت منجر شوند و داده های حیاتی و کاربردی را حذف کنند، چرا که داده های پرت در بعضی مواقع کارایی های بخصوص خود را دارند.

منابع

۱- https://blog.faradars.org/local-outlier-factor-algorithm/
۲- https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-dba۱۶۶۰۸d۴۰۴
۳- https://pyod.readthedocs.io/en/latest/
۴- https://machinelearningmastery.com/model-based-outlier-detection-and-removal-in-python/
۵- /outlier-detection-python-۰۲/۲۰۱۹https://www.analyticsvidhya.com/blog/ pyod/