مقدمه:

در این گزارش می‌خواهیم به سه سوال زیر در خصوص پیش‌پردازش داده‌ها پاسخ دهیم. در هر بخش، سوال مربوطه و در زیر آن پاسخ به آن سوال آورده شده است.

سوالات:

1- دیتاست این سوال با اسم Datapreprocessing در فایل این تمرین ضمیمه شده است. قصد ما در این سوال یادگیری پیش‌پردازش داده‌ها است.

توضیحی در مورد دیتاست: این دیتاست در مورد چند کشور مختلف است که در آن ویژگی‌هایی مانند جمعیت این کشورها، رشد جمعیت و وضعیت توریستی وجود دارد که به کمک این ویژگی‌ها قرار است مدل رگرسیونی بسازیم که تعداد بیماران مبتلا به ویروس کرونا در کشورهای دیگر را تخمین بزنیم. البته دقت کنید با توجه به تعداد کم داده‌ها و اینکه ویژگی‌ها لزوما ویژگی‌های دقیقی نیست، انتظار نتیجه دقیقی نداریم و فقط و فقط هدف یادگیری است.

****

الف) راه‌های مختلف مقابله با Missing Values را به کار ببرید و به نظر شما کدام یک از راه‌ها مناسب‌تر است؟ آیا می‌توان نظر کلی داد؟

یکی از ساده‌ترین و راحت‌ترین راهکارهای رایج در برخورد با Missing Values حذف فیلدها و ویژگی‌هایی است که دارای مقدار خالی هستند. استفاده از این روش عواقبی دارد و باعث از دست رفتن اطلاعاتی می‌شود که ممکن است حائز اهمیت باشد. بخصوص در پایگاه‌داده مورد نظر که تعداد تمام داده‌ها 16 مورد است! بنابراین با حذف مواردی که دارای Missing Values هستند، عملا داده‌های زیادی باقی نمی‌ماند. علاوه بر این، حذف یک داده، به خاطر تنها خالی بودن یک ویژگی آن، کار بیهوده‌ای است! در حقیقت، Schmueli و همکاران در [2] توضیح داده‌اند که اگر فقط 5٪ از مقادیر داده از یک مجموعه‌داده با 30 متغیر گم[1] شوند و مقادیر گمشده به طور مساوی در سراسر مجموعه‌داده پخش شوند، تقریباً 80٪ از داده‌ها حداقل یک مقدار از دست رفته دارند. به همین دلیل تحلیل‌گران داده، ترجیح دادند به سراغ روش‌هایی بروند که در آن مقدار گمشده را با مقدار مناسبی جایگزین کنند. این فرایند جایگزین کردن، دارای ملاک‌ها و معیارهایی است که بعضی از این معیارهای رایج در زیر آمده است:

[1] Miss

- جایگزین کردن مقدار گمشده با مقداری ثابت که توسط تحلیل‌گر داده، مشخص شده است.
- مقدار گمشده را برای متغیرهای عددی با میانگین و برای برای متغیرهای categorical با مد جایگزین کنید.
- مقادیر گمشده را با مقدار تولید شده به طور تصادفی از توزیع مشاهده شده در متغیر مورد نظر جایگزین کنید.
- جایگزین کردن مقادیر گمشده براساس سایر مشخصات و ویژگی‌های نمونه داده.

حال بر روی پایگاه‌داده مورد نظر، داده‌های از دست رفته را با روش‌های معرفی شده در بالا، با مقدار مناسبی جایگزین می‌کنیم. شکل 1-1 پایگاه‌داده مورد نظر را در حالت اولیه خودش نشان می‌دهد.

	Unnamed: 0	DataSource	Unnamed:1	https://data.worldbank.org	Unnamed:3	Unnamed:4	Unnamed: 6	Unnamed
0	0	Data	NaN	2016	NaN	NaN	NaN	NaN
1	1	CountryName	CountryCode	Population growth	Total population	Area (sq. km)	International Visitors	Coronavirus Cases
2	2	Brazil	BRA	0.817555711	207652865	8358140	B	59324
3	3	Switzerland	CHE	1.077221168	8372098	39516	B	29061
4	4	Germany	DEU	1.193866758	82667685	348900	A	156727
5	5	Denmark	DNK	0.834637611	NaN	42262	B	8575
6	6	Spain	ESP	-0.008048086	46443959	500210	A	223759
7	7	France	FRA	0.407491036	66896109	547557	A	161488
8	8	Japan	JPN	-0.115284177	126994511	364560	B	13231
9	9	Greece	GRC	-0.687542545	10746740	128900	C	2506
10	10	Iran	IRN	1.1487886	80277428	1628760	D	90481
11	11	Kuwait	KWT	2.924206194	4052584	NaN	C	3075
12	12	Morocco	MAR	NaN	35276786	446300	C	4047
13	13	Nigeria	NGA	2.619033526	185989640	910770	D	1182
14	14	Qatar	QAT	3.495069918	2569804	11610	B	10287
15	15	Sweden	SWE	NaN	9903122	407310	C	18640
16	16	India	IND	1.148214693	1324171354	2973190	B	26917

شکل 1-1: پایگاه‌داده Datapreprocessing

دو ردیف اول و همچنین ستون اول باید از پایگاه‌داده حذف شوند، زیرا اضافی هستند. شکل 1-2 پایگاه‌داده تمیز شده را نشان می‌دهد.

	CountryName	CountryCode	Population growth	Total population	Area (sq. km)	International Visitors	Coronavirus Cases
0	Brazil	BRA	0.817556	2.076529e+08	8358140.0	B	59324
1	Switzerland	CHE	1.077221	8.372098e+06	39516.0	B	29061
2	Germany	DEU	1.193867	8.266768e+07	348900.0	A	156727
3	Denmark	DNK	0.834638	NaN	42262.0	B	8575
4	Spain	ESP	-0.008048	4.644396e+07	500210.0	A	223759
5	France	FRA	0.407491	6.689611e+07	547557.0	A	161488
6	Japan	JPN	-0.115284	1.269945e+08	364560.0	B	13231
7	Greece	GRC	-0.687543	1.074674e+07	128900.0	C	2506
8	Iran	IRN	1.148789	8.027743e+07	1628760.0	D	90481
9	Kuwait	KWT	2.924206	4.052584e+06	NaN	C	3075
10	Morocco	MAR	NaN	3.527679e+07	446300.0	C	4047
11	Nigeria	NGA	2.619034	1.859896e+08	910770.0	D	1182
12	Qatar	QAT	3.495070	2.569804e+06	11610.0	B	10287
13	Sweden	SWE	NaN	9.903122e+06	407310.0	C	18640
14	India	IND	1.148215	1.324171e+09	2973190.0	B	26917

شکل 1-2: پایگاه‌داده Datapreprocessing با حذف سطرها و ستون‌های اضافه

فیلدهایی که مقدار NaN دارند، داده‌های از دست رفته و یا همان Missing Values هستند. همان‌طور که در شکل 1-2 دیده می‌شود، به طور کلی در این پایگاه‌داده چهار مقدار از دست رفته وجود دارد. شکل زیر تعداد این مقادیر را در هر ستون از داده نشان می‌دهد.

```
 CountryName             0
CountryCode              0
Population growth        2
Total population         1
Area (sq. km)            1
International Visitors   0
Coronavirus Cases        0
dtype: int64
```

اگر بخواهیم ساده‌ترین روش را پیاده کنیم و تمام سطرهایی که دارای Missing Values هستند را حذف کنیم، پایگاه‌داده به صورت شکل 1-3 خواهد شد.

	CountryName	CountryCode	Population growth	Total population	Area (sq. km)	International Visitors	Coronavirus Cases
0	Brazil	BRA	0.817556	2.076529e+08	8358140.0	B	59324
1	Switzerland	CHE	1.077221	8.372098e+06	39516.0	B	29061
2	Germany	DEU	1.193867	8.266768e+07	348900.0	A	156727
4	Spain	ESP	-0.008048	4.644396e+07	500210.0	A	223759
5	France	FRA	0.407491	6.689611e+07	547557.0	A	161488
6	Japan	JPN	-0.115284	1.269945e+08	364560.0	B	13231
7	Greece	GRC	-0.687543	1.074674e+07	128900.0	C	2506
8	Iran	IRN	1.148789	8.027743e+07	1628760.0	D	90481
11	Nigeria	NGA	2.619034	1.859896e+08	910770.0	D	1182
12	Qatar	QAT	3.495070	2.569804e+06	11610.0	B	10287
14	India	IND	1.148215	1.324171e+09	2973190.0	B	26917

شکل 1-3: پایگاه‌داده Datapreprocessing با حذف سطرهای دارای Missing Values

به دلیل اینکه تعداد داده‌ها کم است، روش حذف داده‌ها به دلیل وجود Missing Values کار درستی نیست. روش بعدی این است که توسط تحلیل‌گر عدد ثابتی برای جایگزینی پیشنهاد شود، می‌توان به جای مقادیر عددی مقدار صفر قرار داد. در این پایگاه‌داده، مقادیر غیر عددی که از دست رفته باشد نیز وجود ندارد. شکل 1-4 پایگاه‌داده را با جایگزین شدن مقدار صفر به جای مقادیر Missing Values نشان می‌دهد.

	CountryName	CountryCode	Population growth	Total population	Area (sq. km)	International Visitors	Coronavirus Cases
0	Brazil	BRA	0.817556	2.076529e+08	8358140.0	B	59324
1	Switzerland	CHE	1.077221	8.372098e+06	39516.0	B	29061
2	Germany	DEU	1.193867	8.266768e+07	348900.0	A	156727
3	Denmark	DNK	0.834638	0.000000e+00	42262.0	B	8575
4	Spain	ESP	-0.008048	4.644396e+07	500210.0	A	223759
5	France	FRA	0.407491	6.689611e+07	547557.0	A	161488
6	Japan	JPN	-0.115284	1.269945e+08	364560.0	B	13231
7	Greece	GRC	-0.687543	1.074674e+07	128900.0	C	2506
8	Iran	IRN	1.148789	8.027743e+07	1628760.0	D	90481
9	Kuwait	KWT	2.924206	4.052584e+06	0.0	C	3075
10	Morocco	MAR	0.000000	3.527679e+07	446300.0	C	4047
11	Nigeria	NGA	2.619034	1.859896e+08	910770.0	D	1182
12	Qatar	QAT	3.495070	2.569804e+06	11610.0	B	10287
13	Sweden	SWE	0.000000	9.903122e+06	407310.0	C	18640
14	India	IND	1.148215	1.324171e+09	2973190.0	B	26917

شکل 1-4: پایگاه‌داده Datapreprocessing با جایگزینی مقدار صفر در Missing Values

روش بعدی جایگزین کردن میانگین داده‌های ستون در مقدار از دست رفته است. تعداد مقادیر از دسته رفته چهار مورد است. چون هر چهار مورد، از جمله ویژگی‌های عددی هستند، مقدار از دست رفته را با میانگین اعداد جایگزین می‌کنیم.

```
0     8358140.0
1       39516.0
2      348900.0
3       42262.0
4      500210.0
5      547557.0
6      364560.0
7      128900.0
8     1628760.0
9     1193427.5
10     446300.0
11     910770.0
12      11610.0
13     407310.0
14    2973190.0
Name: Area (sq. km), dtype: float64
```

```
0     0.817556
1     1.077221
2     1.193867
3     0.834638
4    -0.008048
5     0.407491
6    -0.115284
7    -0.687543
8     1.148789
9     2.924206
10    1.142708
11    2.619034
12    3.495070
13    1.142708
14    1.148215
Name: Population growth, dtype: float64
```

```
2.076529e+08
3.372098e+06
3.266768e+07
1.565725e+08
1.644396e+07
3.689611e+07
1.269945e+08
1.074674e+07
3.027743e+07
1.052584e+06
3.527679e+07
1.859896e+08
2.569804e+06
9.903122e+06
1.324171e+09
Total population, dtype: float64
```

شکل 1-5: ستون‌هایی از ویژگی‌هایی که دارای Missing Values هستند با میانگین مقادیر ستون جایگزین شده‌اند.

شکل 1-5 مقدار جایگزین شده میانگین را برای هر ستون از ویژگی نشان می‌دهد.

به طور کلی، روش استفاده از میانگین داده‌ها برای جایگزینی داده از دست رفته، روش قابل قبولی است، اما نباید این حقیقت را فراموش کنیم که با استفاده از این روش، داده‌ای که جایگزین می‌شود، داده‌ی ساختگی است. استفاده از این داده ساختگی، ممکن است در بعضی موارد کارساز باشد، اما کاربر نهایی باید این حقیقت را بداند که داده از دست رفته به این روش جایگزین شده است. با این حال، میانگین ممکن است همیشه بهترین انتخاب برای جایگزینی مقدار نباشد. به عنوان مثال، Larose [3] مجموعه‌داده‌ای را بررسی می‌کند که میانگین آن از 81 درصد داده‌ها بزرگ‌تر است. همچنین، اگر بسیاری از مقادیر گمشده با میانگین جایگزین شوند، سطح

اطمینان نتایج برای استنتاج آماری بیش از حد بدبینانه خواهد شد. در شکل 1-5 نیز، همان‌طور که مشاهده می‌شود میانگین به دست آمده از بیشتر مقادیر موجود در ویژگی مورد نظر بزرگ‌تر است و کمتر شبیه داده‌های واقعی است که وجود دارد. به عنوان مثال در ویژگی Area تنها سه داده دارای عدد 7 رقمی هستند و سایر داده‌ها اعدادی کوچک‌تر از میانگین (سطر شماره 9) دارند. برای اینکه تغییر تمام داده‌های از دست رفته با میانگین جایگزین شود و به طور یکجا نمایش داده شود، ما از SimpleImputer موجود در کتابخانه sklearn.impute استفاده کردیم که در شکل 1-6 نمایش داده شده است. با این روش می‌توان داده‌های categorical را هم به راحتی با مد جایگزین کرد، البته در پایگاه‌داده مورد نظر این نوع داده‌ها که دارای Missing Values باشند وجود ندارد.

	CountryName	CountryCode	Population growth	Total population	Area (sq. km)	International Visitors
0	Brazil	BRA	0.817556	2.076529e+08	8358140.0	B
1	Switzerland	CHE	1.077221	8.372098e+06	39516.0	B
2	Germany	DEU	1.193867	8.266768e+07	348900.0	A
3	Denmark	DNK	0.834638	1.565725e+08	42262.0	B
4	Spain	ESP	-0.008048	4.644396e+07	500210.0	A
5	France	FRA	0.407491	6.689611e+07	547557.0	A
6	Japan	JPN	-0.115284	1.269945e+08	364560.0	B
7	Greece	GRC	-0.687543	1.074674e+07	128900.0	C
8	Iran	IRN	1.148789	8.027743e+07	1628760.0	D
9	Kuwait	KWT	2.924206	4.052584e+06	1193427.5	C
10	Morocco	MAR	1.142708	3.527679e+07	446300.0	C
11	Nigeria	NGA	2.619034	1.859896e+08	910770.0	D
12	Qatar	QAT	3.495070	2.569804e+06	11610.0	B
13	Sweden	SWE	1.142708	9.903122e+06	407310.0	C
14	India	IND	1.148215	1.324171e+09	2973190.0	B

شکل 1-6: پایگاه‌داده Datapreprocessing با جایگزینی مقدار میانگین داده‌های هر ستون در Missing Values

روش بعدی، استفاده از الگوریتم knn برای جایگزین کردن Missing Values است. به عنوان مثال n نزدیک‌ترین همسایه را پیدا می‌کند و سپس از آن‌ها میانگین می‌گیرد و مقدار میانگین را در Missing Values جایگزین می‌کند. شکل 1-7 مقادیر جایگزین شده را توسط این روش نشان می‌دهد.

	CountryName	CountryCode	Population growth	Total population	Area (sq. km)	International Visitors
0	Brazil	BRA	0.817556	2.076529e+08	8358140.0	B
1	Switzerland	CHE	1.077221	8.372098e+06	39516.0	B
2	Germany	DEU	1.193867	8.266768e+07	348900.0	A
3	Denmark	DNK	0.834638	6.212341e+06	42262.0	B
4	Spain	ESP	-0.008048	4.644396e+07	500210.0	A
5	France	FRA	0.407491	6.689611e+07	547557.0	A
6	Japan	JPN	-0.115284	1.269945e+08	364560.0	B
7	Greece	GRC	-0.687543	1.074674e+07	128900.0	C
8	Iran	IRN	1.148789	8.027743e+07	1628760.0	D
9	Kuwait	KWT	2.924206	4.052584e+06	26936.0	C
10	Morocco	MAR	0.413295	3.527679e+07	446300.0	C
11	Nigeria	NGA	2.619034	1.859896e+08	910770.0	D
12	Qatar	QAT	3.495070	2.569804e+06	11610.0	B
13	Sweden	SWE	0.073548	9.903122e+06	407310.0	C
14	India	IND	1.148215	1.324171e+09	2973190.0	B

شکل 1-7: پایگاه‌داده Datapreprocessing با جایگزینی مقدار میانگین n نزدیک‌ترین همسایه در Missing Values

روش استفاده از الگوریتم knn نسبت به سایر روش‌هایی که تاکنون بررسی شد، روش بهتری است زیرا نزدیک-ترین مقادیر را نسبت به داده از دست رفته مورد بررسی پیدا می‌کند و سپس از این مقادیر میانگین می‌گیرد. اما به طور کلی نمی‌توان نظر داد که کدام روش بهترین است، چرا که باید عملیاتی که می‌خواهیم بر روی این پایگاه‌داده اعمال کنیم(مانند رگرسیون) را با حالت‌های مختلف باید انجام دهیم و ببینیم کدام روش بیشترین دقت را به همراه دارد.

****

ب) چرا نمی‌توان از categorical variables برای داده‌کاوی استفاده نمود و باید آن را به ویژگی‌های عددی تبدیل نمود؟ چگونه آن‌ها را می‌توان به متغیرهای عددی تبدیل نمود؟ این روش را بر روی این دیتاست اعمال کنید.

از آن‌جا که بسیاری از الگوریتم‌های یادگیری ماشین نیاز به داده‌های عددی دارند و برای داده‌کاوی باید از این الگوریتم‌ها استفاده کنیم، نیاز است که داده‌های categorical را به داده‌های عددی تبدیل نماییم. تبدیل داده‌های categorical به داده‌های عددی به این صورت است که وقتی داده‌ی categorical داریم، ابتدا به ازای هر مقدار یک مقدار عددی در نظر می‌گیریم و سپس به ازای هر فیلد آن یک ستون یا ویژگی جدید به داده‌ها اضافه می‌کنیم. مثلا در پایگاه‌داده Datapreprocessing ستون International Visitor دارای چهار مقدار A, B, C, D است، به ازای هر کدام از این‌ها یک ستون در نظر می‌گیریم و اگر داده مورد نظر مقدار A دارد، آن را برابر 1 و سایر ستون‌های B, C, D را برابر صفر قرار می‌دهیم. البته می‌توان به جای چهار ستون سه ستون در نظر گرفت، زیرا وقتی هیچ کدام از سه مقدار دیگر نباشد، مقدار چهارم خواهد شد.

پس ابتدا به ویژگی‌های categorical یک مقدار عددی نسبت می‌دهیم. این کار را با استفاده از ماژول sklearn.preprocessing در کتابخانه LabelEncoder انجام می‌دهیم. شکل 1-8 این حالت را نشان می‌دهد.

CountryCode	Population growth	Total population	Area (sq. km)	International Visitors
0	0.817556	2.076529e+08	8358140.0	1
1	1.077221	8.372098e+06	39516.0	1
2	1.193867	8.266768e+07	348900.0	0
3	0.834638	6.212341e+06	42262.0	1
4	-0.008048	4.644396e+07	500210.0	0
5	0.407491	6.689611e+07	547557.0	0
9	-0.115284	1.269945e+08	364560.0	1
6	-0.687543	1.074674e+07	128900.0	2
8	1.148789	8.027743e+07	1628760.0	3
10	2.924206	4.052584e+06	26936.0	2
11	0.413295	3.527679e+07	446300.0	2
12	2.619034	1.859896e+08	910770.0	3
13	3.495070	2.569804e+06	11610.0	1
14	0.073548	9.903122e+06	407310.0	2
7	1.148215	1.324171e+09	2973190.0	1

شکل 1-8: دادن مقدار عددی به ویژگی‌های categorical

حال ستون International Visitors را حذف می‌کنیم و به ازای هر مقدار آن یک ستون در نظر می‌گیریم و مقادیر داده را برای هر کدام در نظر می‌گیریم. شکل 1-9 این حالت را نشان می‌دهد.

CountryCode	Population growth	Total population	Area (sq. km)	A	B	C	D
0	0.817556	2.076529e+08	8358140.0	0.0	1.0	0.0	0.0
1	1.077221	8.372098e+06	39516.0	0.0	1.0	0.0	0.0
2	1.193867	8.266768e+07	348900.0	1.0	0.0	0.0	0.0
3	0.834638	6.212341e+06	42262.0	0.0	1.0	0.0	0.0
4	-0.008048	4.644396e+07	500210.0	1.0	0.0	0.0	0.0
5	0.407491	6.689611e+07	547557.0	1.0	0.0	0.0	0.0
9	-0.115284	1.269945e+08	364560.0	0.0	1.0	0.0	0.0
6	-0.687543	1.074674e+07	128900.0	0.0	0.0	1.0	0.0
8	1.148789	8.027743e+07	1628760.0	0.0	0.0	0.0	1.0
10	2.924206	4.052584e+06	26936.0	0.0	0.0	1.0	0.0
11	0.413295	3.527679e+07	446300.0	0.0	0.0	1.0	0.0
12	2.619034	1.859896e+08	910770.0	0.0	0.0	0.0	1.0
13	3.495070	2.569804e+06	11610.0	0.0	1.0	0.0	0.0
14	0.073548	9.903122e+06	407310.0	0.0	0.0	1.0	0.0
7	1.148215	1.324171e+09	2973190.0	0.0	1.0	0.0	0.0

شکل 1-9: اختصاص دادن یک ستون به ازای هر مقدار در ویژگی categorical

در نهایت شکل 1-10 داده‌ای را نشان می‌دهد که بدون Missing Values و بدون categorical داده است و داده رشته‌ای به عددی تبدیل شده است. برای اینکه همبستگی بین ویژگی‌ها کاهش یابد، داده International Visitors_D را حذف کرده‌ایم تا عمل رگرسیون با دقت بالاتری انجام پذیرد. در نهایت پایگاه‌داده نشان داده شده در شکل زیر، داده مناسب برای انجام عمل رگرسیون است.

	CountryCode	Population growth	Total population	Area (sq. km)	Coronavirus Cases	International Visitors_A	International Visitors_B	International Visitors_C
0	0	0.817556	2.076529e+08	8358140.0	59324	0	1	0
1	1	1.077221	8.372098e+06	39516.0	29061	0	1	0
2	2	1.193867	8.266768e+07	348900.0	156727	1	0	0
3	3	0.834638	6.212341e+06	42262.0	8575	0	1	0
4	4	-0.008048	4.644396e+07	500210.0	223759	1	0	0
5	5	0.407491	6.689611e+07	547557.0	161488	1	0	0
6	9	-0.115284	1.269945e+08	364560.0	13231	0	1	0
7	6	-0.687543	1.074674e+07	128900.0	2506	0	0	1
8	8	1.148789	8.027743e+07	1628760.0	90481	0	0	0
9	10	2.924206	4.052584e+06	26936.0	3075	0	0	1
10	11	0.413295	3.527679e+07	446300.0	4047	0	0	1
11	12	2.619034	1.859896e+08	910770.0	1182	0	0	0
12	13	3.495070	2.569804e+06	11610.0	10287	0	1	0
13	14	0.073548	9.903122e+06	407310.0	18640	0	0	1
14	7	1.148215	1.324171e+09	2973190.0	26917	0	1	0

شکل 1-10: پایگاه‌داده Datapreprocessing پیش‌پردازش شده

****

ج) به نظر شما در این دیتاست آیا به feature scaling نیاز داریم یا خیر؟ چرا؟ اگر پاسخ شما مثبت است، روش-های مختلف feature scaling را بر روی این دیتاست اعمال کنید و از نظر شما کدام مناسب‌تر است؟

بله در این دیتاست به feature scaling نیاز داریم. feature scaling یعنی رنج تغییرات داده‌ها در ویژگی‌های مختلف یکسان باشد. دلیل این امر این است که الگوریتم‌های یادگیری ماشین در صورتی که رنج تغییرات یکسان باشد بهتر عمل می‌کنند. دقت عمل رگرسیون بهتر خواهد بود اگر بر روی ویژگی‌های پایگاه‌داده عمل feature scaling انجام شود.

یکی از روش‌های feature scaling این است که با استفاده از ماژول MinMaxScaler در کتابخانه sklearn.preprocessing به ازای هر ویژگی، هر مقدار را منهای مینیمم هر ویژگی و تقسیم بر رنج هر ویژگی می‌کند. منظور از رنج ویژگی حاصل تفریق ماکزیمم از مینیمم است. با این کار ویژگی‌ها تماما در رنج (0و 1) قرار می‌گیرند.

در پایگاه‌داده مورد نظر، بهتر است سه ویژگی Population growth و Total population و Area (sq. km) در یک بازه مقداردهی شوند. شکل 1-11 این حالت را نشان می‌دهد.

	Population growth	Total population	Area (sq. km)
0	0.359846	0.155178	1.000000
1	0.421929	0.004390	0.003343
2	0.449817	0.060607	0.040411
3	0.363930	0.002756	0.003672
4	0.162457	0.033198	0.058539
5	0.261806	0.048673	0.064212
6	0.136818	0.094147	0.042287
7	0.000000	0.006187	0.014053
8	0.439039	0.058798	0.193751
9	0.863515	0.001122	0.001836
10	0.263194	0.024748	0.052080
11	0.790553	0.138786	0.107729
12	1.000000	0.000000	0.000000
13	0.181965	0.005549	0.047409
14	0.438902	1.000000	0.354828

شکل 1-11: feature scaling به روش MinMaxScaler

روش بعدی feature scaling این است که یک توزیع نرمال از داده‌ها به دست آورد. طوری که میانگین داده‌ها را 0 و واریانس را 1 قرار می‌دهد. این کار را به این صورت انجام می‌دهد که هر ویژگی را منهای میانگین مقادیر آن ویژگی و تقسیم بر انحراف معیار می‌کند. برای سه ویژگی مذکور این حالت در شکل 1-12 نشان داده شده است.

	Population growth	Total population	Area (sq. km)
0	-0.180705	0.190294	3.487210
1	0.047911	-0.430316	-0.518157
2	0.150608	-0.198941	-0.369191
3	-0.165666	-0.437042	-0.516835
4	-0.907588	-0.311750	-0.296336
5	-0.541737	-0.248057	-0.273539
6	-1.002001	-0.060896	-0.361651
7	-1.505832	-0.422920	-0.475120
8	0.110920	-0.206385	0.247054
9	1.674043	-0.443768	-0.524215
10	-0.536627	-0.346528	-0.322293
11	1.405361	0.122830	-0.098654
12	2.176646	-0.448386	-0.531594
13	-0.835749	-0.425548	-0.341067
14	0.110415	3.667412	0.894389

شکل 1-12: feature scaling به روش StandardScaler

روش اول یک سری ایرادات دارد، اینکه داده‌ها را به توزیع نرمال نمی‌برد و یا تاثیر داده‌های پرت را کم نمی‌کند اما روش دوم تقریبا یک توزیع نرمال از داده‌ها به دست می‌آورد و روش دوم برای رگرسیون مناسب‌تر است. روش دوم با استفاده از ماژول StandardScaler در کتابخانه sklearn.preprocessing انجام می‌شود.

روش سوم استفاده از ماژول Normalizer در کتابخانه sklearn.preprocessing است. این روش به این صورت است که مقدار هر سطر را بر نُرم آن سطر تقسیم می‌کند. شکل 1-13 این روش را نشان می‌دهد.

	Population growth	Total population	Area (sq. km)
0	3.933942e-09	0.999191	0.040218
1	1.286666e-07	0.999989	0.004720
2	1.444163e-08	0.999991	0.004220
3	1.343484e-07	0.999977	0.006803
4	-1.732759e-10	0.999942	0.010770
5	6.091197e-09	0.999967	0.008185
6	-9.077849e-10	0.999996	0.002871
7	-6.397225e-08	0.999928	0.011993
8	1.430729e-08	0.999794	0.020285
9	7.215499e-07	0.999978	0.006646
10	1.171483e-08	0.999920	0.012650
11	1.408144e-08	0.999988	0.004897
12	1.360039e-06	0.999990	0.004518
13	7.420428e-09	0.999155	0.041095
14	8.671172e-10	0.999997	0.002245

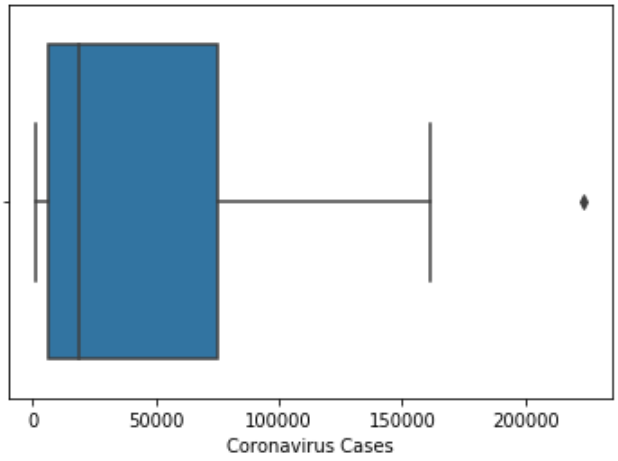
شکل 1-13: feature scaling به روش Normalizer

با توجه به اینکه روش دوم یعنی استفاده از StandardScaler یک توزیع نرمال شده‌ای از داده‌ها به دست می‌آورد، این روش برای feature scaling و بر روی این پایگاه‌داده مناسب‌تر به نظر می‌رسد.

****

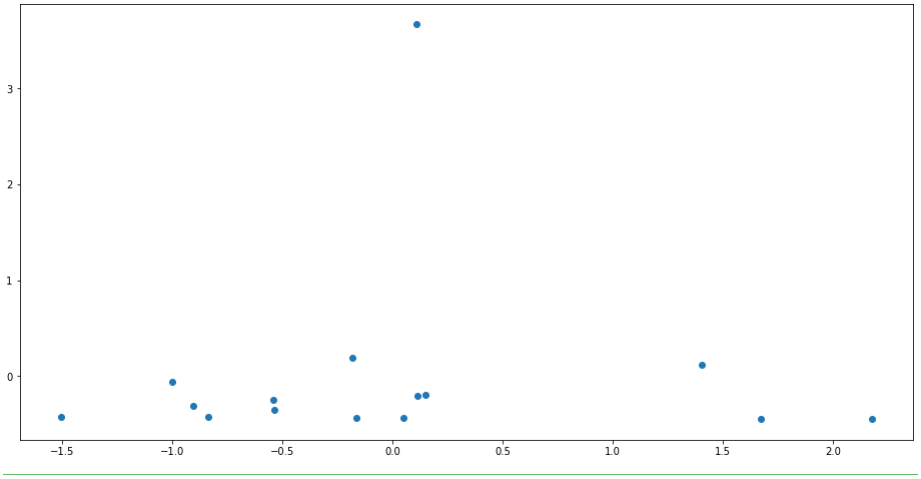
د) داده‌های پرت را در این دیتاست مشخص کنید و آن‌ها را حذف کنید. آیا از نظر شما حذف کردن داده‌های پرت روش درستی است؟ اگر جواب شما منفی است، راه‌حل جایگزین ارائه دهید.

یکی از روش‌ها برا پیدا کردن داده‌های پرت استفاده از نمایش داده‌ها است. به عنوان مثال در ویژگی Coronavirus Cases می‌خواهیم داده‌های پرت را پیدا کنیم، با نمایش داده‌های این ستون داده پرت مشخص می‌شود. شکل 1-14 داده‌های این ستون را نشان می‌دهد. همان طور که در این شکل مشخص است، این ستون یک داده پرت دارد.



شکل 1-14: نمایش داده‌های ستون Coronavirus Cases

روش بعدی این است که یک نمودار نقطه‌ای دو بعدی از ویژگی‌ها رسم کنیم تا داده‌های پرت را پیدا کنیم. به عنوان مثال دو ویژگی Population growth و Total population را نسبت به هم رسم می‌کنیم. شکل 1-15 این نمودار را نشان می‌دهد.



شکل 1-14: نمودار نقطه‌ای دو ویژگی Population growth و Total population

روش بعدی برای پیدا کردن داده‌های پرت این است که از محاسبات آماری و ریاضیات استفاده کنیم. ابتدا z-score را برای این داده‌ها محاسبه می‌کنیم. شکل 1-15 این مقادیر را برای پایگاه‌داده مورد نظر نشان می‌دهد.

	0	1	2	3	4	5	6	7
0	1.620185	0.180705	0.190294	3.487210	0.5	1.224745	0.603023	0.078032
1	1.388730	0.047911	0.430316	0.518157	0.5	1.224745	0.603023	0.361669
2	1.157275	0.150608	0.198941	0.369191	2.0	0.816497	0.603023	1.493232
3	0.925820	0.165666	0.437042	0.516835	0.5	1.224745	0.603023	0.659316
4	0.694365	0.907588	0.311750	0.296336	2.0	0.816497	0.603023	2.467161
5	0.462910	0.541737	0.248057	0.273539	2.0	0.816497	0.603023	1.562406
6	0.462910	1.002001	0.060896	0.361651	0.5	1.224745	0.603023	0.591668
7	0.231455	1.505832	0.422920	0.475120	0.5	0.816497	1.658312	0.747495
8	0.231455	0.110920	0.206385	0.247054	0.5	0.816497	0.603023	0.530722
9	0.694365	1.674043	0.443768	0.524215	0.5	0.816497	1.658312	0.739228
10	0.925820	0.536627	0.346528	0.322293	0.5	0.816497	1.658312	0.725105
11	1.157275	1.405361	0.122830	0.098654	0.5	0.816497	0.603023	0.766732
12	1.388730	2.176646	0.448386	0.531594	0.5	1.224745	0.603023	0.634442
13	1.620185	0.835749	0.425548	0.341067	0.5	0.816497	1.658312	0.513079
14	0.000000	0.110415	3.667412	0.894389	0.5	1.224745	0.603023	0.392820

شکل 1-15: z-score برای پایگاه‌داده Datapreprocessing

حال می‌خواهیم ببینیم در شکل 1-15 کدام مقادیر داده پرت هستند. مقادیر بزرگتر از 3 داده پرت هستند. در این پایگاه‌داده دو مقدار data[0][3] و data[14][2] داده پرت هستند.

می‌توانیم داده‌های پرت را حذف کنیم، اما این روش، روش مناسبی نیست، روش بهتر می‌تواند این باشد که داده پرت را تغییر دهیم. به عنوان مثال می‌توان داده پرت را در همان ویژگی که outlier محسوب شده، به عنوان Missing Value در نظر بگیریم و بر اساس روش‌های جایگزین کردن مقدار در Missing Value آن را تغییر دهیم تا از حالت outlier خارج شود. شکل 1-16 حالتی را نشان می‌دهد داده‌های پرت را حذف کردیم.

	CountryCode	Population growth	Total population	Area (sq. km)	International Visitors_A	International Visitors_B	International Visitors_C	Coronavirus Cases
1	1	0.047911	-0.430316	-0.518157	0	1	0	29061
2	2	0.150608	-0.198941	-0.369191	1	0	0	156727
3	3	-0.165666	-0.437042	-0.516835	0	1	0	8575
4	4	-0.907588	-0.311750	-0.296336	1	0	0	223759
5	5	-0.541737	-0.248057	-0.273539	1	0	0	161488
6	9	-1.002001	-0.060896	-0.361651	0	1	0	13231
7	6	-1.505832	-0.422920	-0.475120	0	0	1	2506
8	8	0.110920	-0.206385	0.247054	0	0	0	90481
9	10	1.674043	-0.443768	-0.524215	0	0	1	3075
10	11	-0.536627	-0.346528	-0.322293	0	0	1	4047
11	12	1.405361	0.122830	-0.098654	0	0	0	1182
12	13	2.176646	-0.448386	-0.531594	0	1	0	10287
13	14	-0.835749	-0.425548	-0.341067	0	0	1	18640

شکل 1-16: پایگاه‌داده پیش‌پردازش شده Datapreprocessing با حذف داده‌های پرت

حال به جای داده شماره 0 در ویژگی شماره 3 آن و همین‌طور به جای داده شماره 14 در ویژگی شماره 2 آن مقدار NaN را قرار می‌دهیم و با آن مانند داده از دست رفته برخورد می‌کنیم.

****

۵) فرض کنید تمام داده‌های ما، داده‌های آموزش است. Multiple linear Regression را بر روی این دیتاست پیش‌پردازش شده اعمال کنید. مدل را به طور کامل گزارش کنید. منظور از گزارش مدل، مشخص کردن ضرایب، عرض از مبدا، خطا MSE، RMSE و خروجی به کمک Satesmodels می‌باشد. گزارش خود را تحلیل کنید.

برای Multiple linear Regression ابتدا x و y را تعیین می‌کنیم. شکل 1-17 x را نشان می‌دهد و ستون آخر (Coronavirus Cases) y را نشان می‌دهد.

	CountryCode	Population growth	Total population	Area (sq. km)	International Visitors_A	International Visitors_B	International Visitors_C
1	1	0.047911	-0.430316	-0.518157	0	1	0
2	2	0.150608	-0.198941	-0.369191	1	0	0
3	3	-0.165666	-0.437042	-0.516835	0	1	0
4	4	-0.907588	-0.311750	-0.296336	1	0	0
5	5	-0.541737	-0.248057	-0.273539	1	0	0
6	9	-1.002001	-0.060896	-0.361651	0	1	0
7	6	-1.505832	-0.422920	-0.475120	0	0	1
8	8	0.110920	-0.206385	0.247054	0	0	0
9	10	1.674043	-0.443768	-0.524215	0	0	1
10	11	-0.536627	-0.346528	-0.322293	0	0	1
11	12	1.405361	0.122830	-0.098654	0	0	0
12	13	2.176646	-0.448386	-0.531594	0	1	0
13	14	-0.835749	-0.425548	-0.341067	0	0	1

شکل 1-17: مقادیر x برای رگرسیون خطی

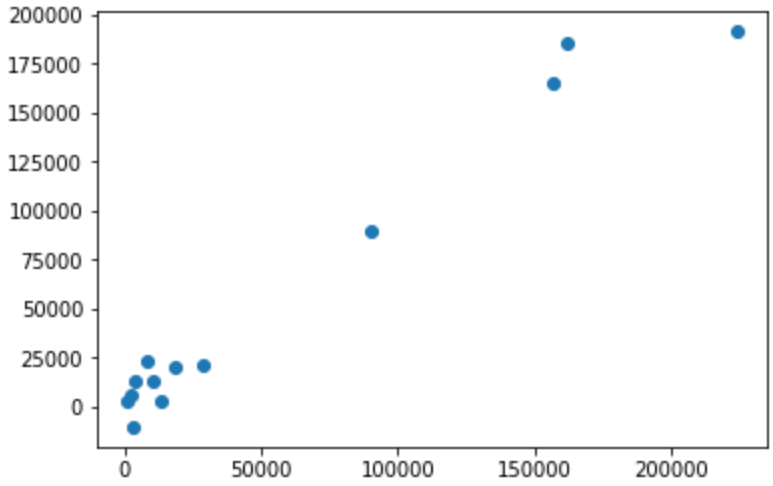
ابتدا مدل را با تمام داده‌ها آموزش می‌دهیم. پس از آموزش عرض از مبدا مقدار 34810.7253468937 است و ضرایب به صورت شکل زیر هستند:

```
array([-1.46443594e+01, -3.90994143e+03, -1.14065704e+05,  1.26228252e+05,
        1.54871492e+05,  3.19875291e+03, -2.30325997e+04])
```

شکل زیر مقادیر ضرایب را برای هر ویژگی به طور جداگانه نشان می‌دهد:

	Coefficient
CountryCode	-14.644359
Population growth	-3909.941425
Total population	-114065.704103
Area (sq. km)	126228.251759
International Visitors_A	154871.492211
International Visitors_B	3198.752906
International Visitors_C	-23032.599688

حال مدل را predict می‌کنیم. شکل 1-18 نمودار نقطه‌ای بین y اصلی و مقدار y ای که با استفاده از رگرسیون خطی تخمین زده شده است را نشان می‌دهد.



شکل 1-18: scatter plot برای مقدار پیش‌بینی شده و مقدار واقعی مدل آموزش دیده با رگرسیون خطی

اگر نمودار شکل 1-18 نزدیک به یک شکل یک خط باشد، نشان‌دهنده این است که مدل به خوبی آموزش دیده است، هر چند در این پایگاه‌داده تعداد داده‌ها اندک است و مدل هم به خوبی آموزش ندیده است.

خطای MAE برای این داده‌های آموزشی مقدار 10028.200328336952 است. این خطا میانگین خطا را به ما نشان می‌دهد و همان‌طور که دیده می‌شود عدد بسیار بالایی است. مقدار خطای MSE در این مدل آموزش دیده شده مقدار 181360724.8434098 است. مقدار خطای RMSE که معیار ملموس‌تری است نیز مقدار 3.2891712515727004e+16 یا 24.940894443839966 است. دلیل این امر می‌توان این باشد که scale ستون y نسبت به مقادیر دیگر بالا است.

دلیل اینکه خطا مقدار قابل قبولی نیست این است که تعداد داده‌ها بسیار کم است و هم اینکه مقادیر ستون خروجی بسیار متفاوت با مقادیر ویژگی‌ها است. در واقع با 13 داده نمی‌توان مدل رگرسیون را به خوبی آموزش داد.

حال با کمک پکیج Satesmodels می‌توانیم فرضیه‌های آماری انجام دهیم و مدل را تحلیل کنیم. با استفاده از این پکیج رگرسیون را می‌سازیم و آموزش می‌دهیم. شکل زیر پارامترهای مدل ساخته شده را نشان می‌دهد.

```
Intercept                      34810.725347
Q("CountryCode")                 -14.644359
Q("Population growth")         -3909.941425
Q("Total population")        -114065.704103
Q("Area (sq. km)")            126228.251759
Q("International Visitors_A")  154871.492211
Q("International Visitors_B")    3198.752906
Q("International Visitors_C")  -23032.599688
dtype: float64
```

شکل‌های زیر مقادیر تحلیل‌های آماری مدل را به طور خلاصه نشان می‌دهد.

Dep. Variable:	Q("Coronavirus Cases")	R-squared:	0.966
Model:	OLS	Adj. R-squared:	0.919
Method:	Least Squares	F-statistic:	20.57
Date:	Thu, 11 Jun 2020	Prob (F-statistic):	0.00211
Time:	20:01:48	Log-Likelihood:	-142.05
No. Observations:	13	AIC:	300.1
Df Residuals:	5	BIC:	304.6
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.481e+04	2.62e+04	1.328	0.242	-3.26e+04	1.02e+05
Q("CountryCode")	-14.6444	2658.317	-0.006	0.996	-6848.065	6818.776
Q("Population growth")	-3909.9414	1.06e+04	-0.369	0.727	-3.12e+04	2.33e+04
Q("Total population")	-1.141e+05	6.49e+04	-1.758	0.139	-2.81e+05	5.28e+04
Q("Area (sq. km)")	1.262e+05	9.76e+04	1.293	0.252	-1.25e+05	3.77e+05
Q("International Visitors_A")	1.549e+05	5.05e+04	3.065	0.028	2.5e+04	2.85e+05
Q("International Visitors_B")	3198.7529	6.67e+04	0.048	0.964	-1.68e+05	1.75e+05
Q("International Visitors_C")	-2.303e+04	7.42e+04	-0.311	0.769	-2.14e+05	1.68e+05

Omnibus:	3.151	Durbin-Watson:	3.288
Prob(Omnibus):	0.207	Jarque-Bera (JB):	1.044
Skew:	0.624	Prob(JB):	0.593
Kurtosis:	3.608	Cond. No.	216.

شکل 1-19: مقادیر آماری با استفاده از پکیج Satesmodels برای مدل آموزش دیده

همان‌طور که در شکل‌های بالا نشان داده شده است، مدل به خوبی آموزش ندیده است و مقادیر p_value برای هیچ کدام از ویژگی‌ها صفر نیست.

برای رفع این مشکل می‌توان از Ridge Regression استفاده کرد. این روش باعث می‌شود مدل بهتر آموزش ببیند و از underfit و overfit شدن جلوگیری شود، هر چند که در این پایگاه‌داده ایراد اصلی در کم بودن تعداد داده است. وقتی مقدار آلفا را 1000 قرار می‌دهیم، خطای RMSE مدل کاهش به مقدار 2.3980213150049382e+19 کاهش می‌یابد.

****

و) فرض کنید ستون Total papulation را در متغیر X قرار دهیم و ستون تعداد مبتلایان به ویروس کرونا را در متغیر y قرار دهیم. حال می‌خواهیم مدل رگرسیونی به فرم $y = ax^2 + bx + c$ بسازیم. پارامترهای مجهول را به دست آورید.

برای اینکه بتوانیم مدل رگرسیونی به این صورت بسازیم باید یک ستون به داده اضافه کنیم و مقدار x^2 را در آن به عنوان یک ویژگی قرار دهیم و سپس مانند بخش قبل از Multiple linear Regression استفاده کنیم. شکل 1-20 مقدار x را نشان می‌دهد.

	x^2	x
0	4.311971e+16	2.076529e+08
1	7.009202e+13	8.372098e+06
2	6.833946e+15	8.266768e+07
3	3.859318e+13	6.212341e+06
4	2.157041e+15	4.644396e+07
5	4.475089e+15	6.689611e+07
6	1.612761e+16	1.269945e+08
7	1.154924e+14	1.074674e+07
8	6.444465e+15	8.027743e+07
9	1.642344e+13	4.052584e+06
10	1.244452e+15	3.527679e+07
11	3.459215e+16	1.859896e+08
12	6.603893e+12	2.569804e+06
13	9.807183e+13	9.903122e+06
14	1.753430e+18	1.324171e+09

شکل 1-20: مقدار x برای ایجاد مدل رگرسیون

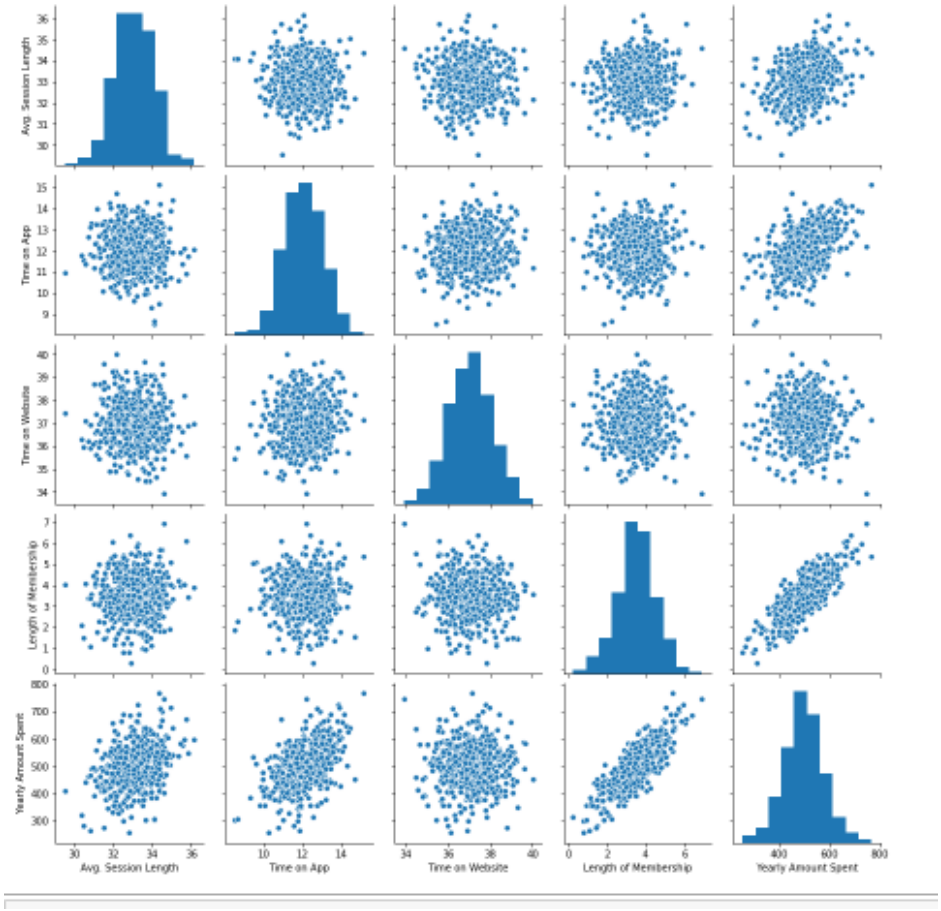
با آموزش دادن مدل، پارامترهای مجهول به این صورت است که عرض از مبدا مقدار 45848.57233761 دارد و دو پارامتر بعدی مقادیر1.49139258e-13- و1.82091435e-04 دارد.

****

2- دیتاست این سوال به اسم Ecommerce Customers است.

الف) ابتدا به کمک pairplot و heatmap در مورد همبستگی ستون‌ها نظر دهید.

شکل 2-1 نمودار ارتباط بین ویژگی‌ها را با استفاده از pairplot نشان می‌دهد.



شکل 2-1: نمودار ارتباط بین ویژگی‌ها با استفاده از pairplot

در شکل 2-1، علاوه بر ارتباط دو به دوی تمام ویژگی‌ها، هیستوگرام آن‌ها نیز نشان داده شده است. هیستوگرام نشان می‌دهد که توزیع تمام داده‌ها گوسی است و دارای توزیع نرمالی است. ویژگی Length of Membership و Yearly Amount Spent دارای ارتباط خطی هستند، همین‌طور دو ویژگی Yearly Amount Spent و Time on App نیز تقریبا با همدیگر ارتباط خطی دارند. حدس ارتباط بین باقی ویژگی‌ها با استفاده از این نمودار دشوار است. برای اینکه همبستگی بین ویژگی‌ها را بررسی کنیم از heatmap استفاده می‌کنیم. شکل 2-2 این نقشه همبستگی بین ویژگی‌ها را نشان می‌دهد.

شکل 2-2: همبستگی بین ویژگی‌ها با استفاده از heatmap

شکل 2-2 به خوبی ارتباط بین ویژگی‌ها را نشان می‌دهد. در این شکل هر چه مربع نشان داده شده بین دو ویژگی کمرنگ‌تر باشد یعنی همبستگی بین آن دو ویژگی بیشتر است. برای عمل رگرسیون و یا طبقه‌بندی هر چه همبستگی بین ویژگی‌ها بیشتر باشد، نتیجه‌ای که به دست خواهیم آورد نامطلوب‌تر خواهد شد. در این پایگاه‌داده دو ویژگی Yearly Amount Spent و Length of Membership بیشترین ارتباط را با همدیگر دارند. اگر بخواهیم نتیجه مطلوبی از رگرسیون بگیریم باید یکی از این ویژگی‌ها را حذف کنیم. پس از این دو ویژگی، ویژگی‌های Yearly Amount Spent و Time on App نیز با یکدیگر همبستگی دارند. با توجه به راهنمای نقشه‌ای که در سمت راست شکل 2-2 وجود دارد، این همبستگی تقریبا به میزان 0.6 است. پس از این ویژگی‌ها، دو ویژگی Avg. Session Length و Yearly Amount Spent نیز حدودا به میزان 0.4 با یکدیگر همبستگی دارند. سایر ویژگی‌ها همبستگی‌ای بین‌شان نیست و نسبت به هم مستقل هستند. ویژگی‌هایی که از هم مستقل‌اند بهترین ویژگی‌ها برای پیاده‌سازی الگوریتم‌های یادگیری ماشین هستند. اما در این پایگاه‌داده تمام ویژگی‌ها به جز Time on website با ویژگی Yearly Amount Spent که target است وابستگی دارند.

****

ب) Multiple linear Regression را بر روی این دیتاست اعمال کنید. مدل را به طور کامل گزارش کنید. منظور از گزارش مدل، مشخص کردن ضرایب، عرض از مبدا، خطا MSE، RMSE داده آموزش و خروجی به کمک Satesmodels می‌باشد. گزارش خود را تحلیل کنید.

حال می‌خواهیم رگرسیون خطی را بر روی این پایگاه‌داده اعمال کنیم. دو ویژگی Email و Address که مشخصات فردی اشخاص است، تاثیری بر روی رگرسیون ندارد. بنابراین قبل از انجام عمل رگرسیون این دو ویژگی را از

پایگاه‌داده حذف می‌کنیم. ویژگی Avatar که رنگ را نشان می‌دهد را نیز که داده رشته‌ای است را به داده عددی تبدیل می‌کنیم. 30 درصد از داده‌ها را برای تست و مابقی را برای آموزش در نظر می‌گیریم.

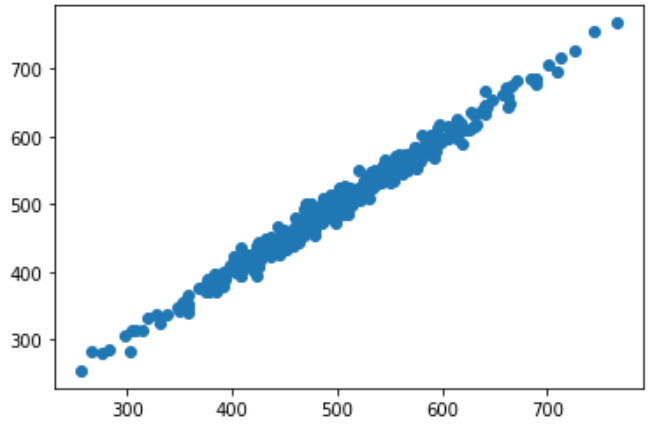
حال که ویژگی‌ها برای رگرسیون آماده شدند، با استفاده از این پایگاه‌داده مدل را آموزش می‌دهیم. عرض از مبدا مقدار 1062.3045544677686- دارد و ضرایب هر ویژگی به صورت زیر هستند:

	Coefficient
Avatar	0.018012
Avg. Session Length	25.985512
Time on App	38.830466
Time on Website	0.400377
Length of Membership	61.818765

شکل زیر خطاهای داده آموزش را نشان می‌دهد:

```
MAE:  7.889980839821057
MSE:  98.72940800266184
RMSE:  9747.496004556067
```

شکل 2-3 نیز نمودار نقطه‌ای بین مقدار واقعی تارگت و خروجی تخمین‌زده شده با استفاده از رگرسیون را نشان می‌دهد.



شکل 2-3: نمودار نقطه‌ای داده‌های تخمین زده شده با استفاه از رگرسیون

همان‌طور که در شکل بالا دیده می‌شود داده‌های آموزش با استفاده از این مدل به خوبی آموزش دیده‌اند.

حال خروجی را براساس Satesmodels بررسی می‌کنیم. براساس این مدل ضرایب به صورت زیر است:

```
Intercept                   -1051.809978
Q("Length of Membership")      61.573135
Q("Time on Website")            0.434799
Q("Time on App")               38.715280
Q("Avg. Session Length")       25.724551
Q("Avatar")                     0.007999
dtype: float64
```

خلاصه‌ای از محاسبات و معیارهای آماری مدل نیز به صورت زیر است:

Dep. Variable:	Q("Yearly Amount Spent")	**R-squared:**	0.984
Model:	OLS	**Adj. R-squared:**	0.984
Method:	Least Squares	**F-statistic:**	6207.
Date:	Fri, 12 Jun 2020	**Prob (F-statistic):**	0.00
Time:	16:06:21	**Log-Likelihood:**	-1856.6
No. Observations:	500	**AIC:**	3725.
Df Residuals:	494	**BIC:**	3751.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1051.8100	23.005	-45.721	0.000	-1097.010	-1006.610
Q("Length of Membership")	61.5731	0.449	137.262	0.000	60.692	62.454
Q("Time on Website")	0.4348	0.444	0.979	0.328	-0.438	1.308
Q("Time on App")	38.7153	0.451	85.787	0.000	37.829	39.602
Q("Avg. Session Length")	25.7246	0.451	56.984	0.000	24.838	26.612
Q("Avatar")	0.0080	0.011	0.734	0.463	-0.013	0.029

Omnibus:	0.424	**Durbin-Watson:**	1.890
Prob(Omnibus):	0.809	**Jarque-Bera (JB):**	0.275
Skew:	-0.035	**Prob(JB):**	0.872
Kurtosis:	3.091	**Cond. No.**	4.76e+03

شکل 2-4: مقادیر آماری با استفاده از پکیج Satesmodels برای مدل آموزش دیده

****

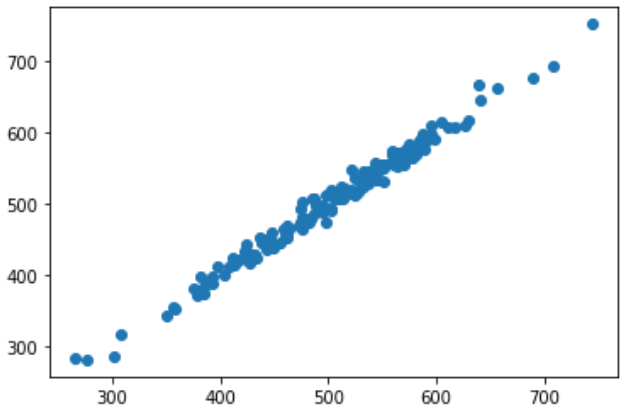
ج) به کمک k-fold cross validation خطای تست را تقریب بزنید.

خطای تست با کمک k-fold cross validation مقدار 100.54600314953888 به دست آمد. با توجه به اینکه این خطا به خطای آموزش که در بخش قبل بیان شد و حدودا مقدار 98.72 داشت، نزدیک است، خطای قابل قبولی است. لازم به ذکر است که برای محاسبه خطا از mean square error استفاده کردیم.

****

د) حال با استفاده از مدل به دست آمده داده‌های تست را پیش‌بینی کنید و RMSE تست را به دست آورید. با توجه به خطای تست و آموزش آیا overfit یا underfit رخ داده است. نظرتان را در مورد bias و variance این مدل بیان کنید. اگر مدل به خوبی کار نکرده است، دلایل آن را بیان کنید.

حال مدل به دست آمده را بر روی داده‌های تست پیش‌بینی می‌کنیم. 30 درصد داده‌ها را به عنوان داده تست در نظر می‌گیریم. شکل 2-5 نمودار نقطه‌ای برای داده‌های تست پیش‌بینی شده و واقعی را نشان می‌دهد.



شکل 2-5: نمودار نقطه‌ای داده‌های تخمین زده شده با استفاده از رگرسیون

خطای داده‌های تست نیز به صورت زیر است:

```
MAE:  7.822865005130022
MSE:  94.85987719874008
RMSE:  8998.396302160048
```

خطای تست از خطای آموزش پایین‌تر است. با اینکه در نمودارهای نقطه‌ای مربوط به داده‌های تست و آموزش شکل نشان می‌دهد که مدل به خوبی آموزش دیده است اما خطا این گمان را ایجاد می‌کند که مدل overfit شده باشد.

****

ه) حال با توجه به ارزیابی خود از مدل به سوالی که شرکت از شما دارد، پاسخ دهید.

با توجه به مدل ساخته شده و با توجه به نقشه همبستگی بین ویژگی‌ها می‌توان به این سوال پاسخ داد که بهتر است شرکت بیشتر بر روی app mobile سرمایه‌گذاری کند و تمرکز بیشتری بر روی آن داشته باشد. چرا که با استفاده از تحلیل مدل و داده‌ها، به این نتیجه می‌رسیم که هر شخصی که میزان استفاده از app بیشتر بوده به مراتب استفاده سالیانه‌اش از محصولات این شرکت نیز بیشتر بوده است در صورتی که میزان استفاده از web بر روی استفاده سالیانه از محصولات شرکت تاثیری ندارد. زیرا این متغیر یک متغیر مستقل از ویژگیYearly Amount Spent است و استفاده از آن بر روی استفاده سالانه تاثیری ندارد.

****

****

3- در این قسمت باید در خروجی همه سوالات معیارهای ROC, Confusion Matrix, Classification Report بیان شود.

الف) دیتاست Arrhythmia دارای Missing value است. با یکی از راه‌های مقابله به داده گمشده، پیش‌پردازش داده را انجام دهید.

با استفاده از الگوریتم knn دو نزدیک‌ترین همسایه را برای هر مقدار گمشده، پیدا می‌کنیم و میانگین آن‌ها را در مکان داده گمشده قرار می‌دهیم.

****

ب) الگوریتم نزدیک‌ترین همسایه را برای k = 1 و k = 30 با معیار فاصله اقلیدسی، بر روی دیتاست دانلود شده اعمال کنید. تمام معیارهای آموزش را برای این داده‌های آموزش و تست به دست آورید و آن‌ها را تحلیل کنید.

بر روی این پایگاه‌داده ابتدا knn را با k = 1 و معیار فاصله اقلیدسی به دست می‌آوریم. 30 درصد از کل داده‌ها را برای داده‌های تست در نظر می‌گیریم. داده‌ها را با استفاده از StandardScaler نرمال می‌کنیم تا نتیجه بهتری به دست آوریم.

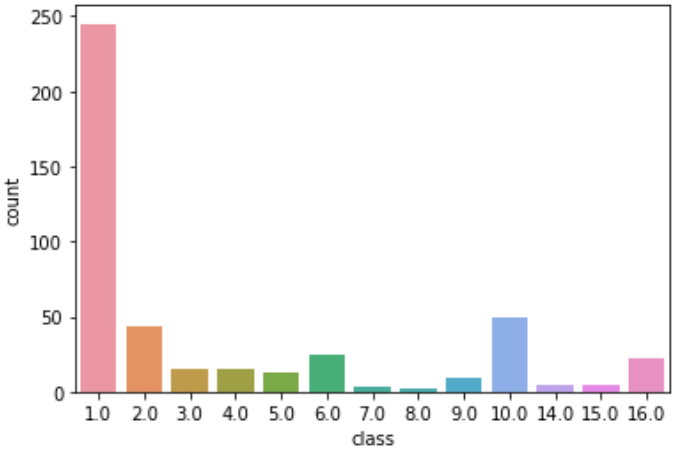
معیارهای زیر برای داده‌های تست است:

```
[[62  0  0  0  2  7  0  0  2  0  1]
 [ 8  1  0  0  0  1  0  0  0  0  0]
 [ 0  0  2  0  0  0  0  0  0  0  0]
 [ 2  0  0  1  0  0  0  0  0  0  0]
 [ 3  1  0  0  0  0  0  0  0  0  1]
 [ 7  0  0  0  0  2  0  0  0  0  1]
 [ 1  0  0  0  0  0  0  0  0  0  0]
 [ 1  1  0  0  0  0  0  1  0  0  0]
 [ 8  0  0  0  1  1  0  0  3  0  0]
 [ 3  1  0  0  0  0  0  0  0  0  0]
 [ 8  2  0  0  1  0  0  0  0  0  0]]
```

شکل 3-1: Confusion Matrix برای داده‌های تست با k=1

تحلیل Confusion Matrix:

شکل 3-1 ماتریس Confusion برای داده‌های تست را نشان می‌دهد. اولا در داده‌های تست از تمام کلاس‌ها وجود ندارد. در صورتی که در داده‌های آموزشی از هر 13 کلاس وجود دارد. برای اینکه تعداد کلاس‌ها و فراوانی آن‌ها را در این پایگاه‌داده بهتر درک کنیم، شکل 3-2 را مشاهده کنید. شکل 3-2 نشان می‌دهد که 16 کلاس مختلف در این پایگاه‌داده وجود دارد که اکثر داده‌ها در کلاس اول هستند. در واقع شکل 3-2 تنوع داده‌ها در کلاس‌های مختلف را نشان می‌دهد.



شکل 3-2: تنوع فراوانی تمام داده‌ها در کلاس‌های مختلف

همان‌طور که در شکل 3-2 نشان داده شده است، تعداد بسیار زیادی از داده‌ها در کلاس 1 قرار دارند. به همین علت در ماتریس Confusion داده‌های تست نیز اکثر داده‌ها جزو این کلاس در نظر گرفته شده اند.

در Confusion Matrix هر چه اعداد موجود در قطر اصلی ماتریس بیشتر باشند، این مفهوم را می‌رسانند که داده‌ها به درستی دسته‌بندی شده‌اند. به عنوان مثال درایه (1و 1) نشان می‌دهد چه تعداد از داده‌ها در کلاس 1 بوده‌اند و به درستی در کلاس 1 دسته‌بندی شده‌اند.

شکل 3-3 Classification Report را برای داده‌های تست نشان می‌دهد.

	precision	recall	f1-score	support
1.0	0.60	0.84	0.70	74
2.0	0.17	0.10	0.12	10
3.0	1.00	1.00	1.00	2
4.0	1.00	0.33	0.50	3
5.0	0.00	0.00	0.00	5
6.0	0.18	0.20	0.19	10
7.0	0.00	0.00	0.00	1
9.0	1.00	0.33	0.50	3
10.0	0.60	0.23	0.33	13
15.0	0.00	0.00	0.00	4
16.0	0.00	0.00	0.00	11
accuracy			0.53	136
macro avg	0.41	0.28	0.30	136
weighted avg	0.47	0.53	0.47	136

شکل 3-3: Classification Report برای داده‌های تست با k=1

با استفاده از شکل 3-3 متوجه می‌شویم که در داده‌های تست از کلاس 8 و 14 داده‌ای وجود ندارد. بررسی مقدار recall برای تک تک کلاس‌ها نشان می‌دهد که چه تعداد از داده‌ها به درستی در کلاس مربوطه دسته‌بندی شده‌اند. همان‌طور که دیده می‌شود برای کلاس 3 تمام داده‌های تست که البته با توجه به Confusion Matrix تنها 2 داده بوده است به درستی توسط knn دسته‌بندی شده‌اند.

ستون f1_score میانگین هارمونیکی از precision و recall است. اگر این معیار برای تمام داده‌ها یک باشد یعنی عمل دسته‌بندی به درستی انجام شده است. در شکل 3-3 می‌بینیم که برای داده‌ها کلاس 3 و 1 عدد بالایی است و برای سایر کلاس‌ها عدد پایینی دارد.

accuracy نیز نشان می‌دهد که چه میزان از داده‌ها را به درستی دسته‌بندی کرده‌ایم. شکل 3-3 نشان می‌دهد که برای داده‌های تست دقت 0.53 است که دقت خوبی نیست. دلیل این امر این است که مقدار k را در الگوریتم knn یک قرار داده‌ایم. هر چند دقت معیار مناسبی نیست، چرا که به بالانس بودن داده‌ها در کلاس‌های مختلف بستگی دارد و در این پایگاه‌داده داده‌ها به مساوات در کلاس‌های مختلف وجود ندارد.

معیارهای زیر برای داده‌های آموزش است:

```
[[171   0   0   0   0   0   0   0   0   0   0   0   0]
 [  0  34   0   0   0   0   0   0   0   0   0   0   0]
 [  0   0  13   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0  12   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   8   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0  15   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   2   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   2   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   6   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0  37   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   4   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   1   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0  11]]
```

شکل 3-4: Confusion Matrix برای داده‌های آموزش با k=1

	precision	recall	f1-score	support
1.0	1.00	1.00	1.00	171
2.0	1.00	1.00	1.00	34
3.0	1.00	1.00	1.00	13
4.0	1.00	1.00	1.00	12
5.0	1.00	1.00	1.00	8
6.0	1.00	1.00	1.00	15
7.0	1.00	1.00	1.00	2
8.0	1.00	1.00	1.00	2
9.0	1.00	1.00	1.00	6
10.0	1.00	1.00	1.00	37
14.0	1.00	1.00	1.00	4
15.0	1.00	1.00	1.00	1
16.0	1.00	1.00	1.00	11
accuracy			1.00	316
macro avg	1.00	1.00	1.00	316
weighted avg	1.00	1.00	1.00	316

شکل 3-5: Classification Report برای داده‌های آموزش با k=1

شکل 3-4 و 3-5 نشان می‌دهد که دقت داده‌های آموزشی 100 درصد شده است و تمام داده‌ها به درستی در کلاس‌های مربوط به خودشان دسته‌بندی شده‌اند! این حالت نشان‌دهنده این است که مدل overfit شده است.

حال بر روی این پایگاه‌داده knn را با k = 30 و معیار فاصله اقلیدسی به دست می‌آوریم:

شکل 3-6 Confusion Matrix را برای داده‌های تست نشان می‌دهد.

```
[[82  0  0  0  0  0  0  0  0  0  0]
 [12  0  0  0  0  0  0  0  0  0  0]
 [ 5  0  0  0  0  0  0  0  0  0  0]
 [ 4  0  0  0  0  0  0  0  0  0  0]
 [ 1  0  0  0  0  0  0  0  0  0  0]
 [ 7  0  0  0  0  0  0  0  0  0  0]
 [ 2  0  0  0  0  0  0  0  0  0  0]
 [14  0  0  0  0  0  0  0  0  0  0]
 [ 1  0  0  0  0  0  0  0  0  0  0]
 [ 1  0  0  0  0  0  0  0  0  0  0]
 [ 7  0  0  0  0  0  0  0  0  0  0]]
```

شکل 3-6: Confusion Matrix برای داده‌های تست با k=30

شکل 3-6 نشان می‌دهد که تمام داده‌ها جزو کلاس 1 دسته‌بندی شده‌اند. دلیل این امر این است که شعاع همسایگی را بسیار زیاد در نظر گرفته‌ایم و از آنجایی که در نمودار میله‌ای شکل 3-2 دیدیم تعداد بسیار زیادی از داده‌ها دارای برچسب 1 هستند. به همین دلیل تمام 136 داده تست برچسب 1 گرفته‌اند.

شکل 3-7 Classification Report را برای داده‌های تست نشان می‌دهد.

	precision	recall	f1-score	support
1.0	0.60	1.00	0.75	82
2.0	0.00	0.00	0.00	12
3.0	0.00	0.00	0.00	5
4.0	0.00	0.00	0.00	4
5.0	0.00	0.00	0.00	1
6.0	0.00	0.00	0.00	7
8.0	0.00	0.00	0.00	2
10.0	0.00	0.00	0.00	14
14.0	0.00	0.00	0.00	1
15.0	0.00	0.00	0.00	1
16.0	0.00	0.00	0.00	7
accuracy			0.60	136
macro avg	0.05	0.09	0.07	136
weighted avg	0.36	0.60	0.45	136

شکل 3-7: Classification Report برای داده‌های تست با k=30

شکل 3-7 نشان می‌دهد که با اینکه تنها برای کلاس 1 مقادیر f1_score و precision و recall دارای مقدار بالایی هستند اما معیار دقت 60 درصد به دست آمده است! زیرا همان‌طور که گفته شد، دقت زمانی کاربردی است که داده‌ها بالانس باشند و در اینجا چون داده‌ها بالانس نیستند و اکثر داده‌ها در کلاس 1 هستند بنابراین این داده‌ها از اهمیت بالایی برای دقت برخوردار خواهند بود که به همین دلیل دقت بالا نشان داده شده است. به عبارت دیگر چون تمام داده‌های کلاس 1 به درستی دسته‌بندی شده‌اند و داده‌های کلاس 1 بیشتر داده‌ها از پایگاه‌داده را تشکیل می‌دهند، دقت بالاتر از حالت قبل که k=1 بود به دست آمد.

شکل 3-8 و 3-9 نیز Confusion Matrix و Classification Report برای داده‌های آموزشی نشان می‌دهند.

```
[[163   0   0   0   0   0   0   0   0   0   0   0]
 [ 32   0   0   0   0   0   0   0   0   0   0   0]
 [ 10   0   0   0   0   0   0   0   0   0   0   0]
 [ 11   0   0   0   0   0   0   0   0   0   0   0]
 [ 12   0   0   0   0   0   0   0   0   0   0   0]
 [ 18   0   0   0   0   0   0   0   0   0   0   0]
 [  3   0   0   0   0   0   0   0   0   0   0   0]
 [  9   0   0   0   0   0   0   0   0   0   0   0]
 [ 35   0   0   0   0   0   0   0   1   0   0   0]
 [  3   0   0   0   0   0   0   0   0   0   0   0]
 [  4   0   0   0   0   0   0   0   0   0   0   0]
 [ 15   0   0   0   0   0   0   0   0   0   0   0]]
```

شکل 3-8: Confusion Matrix برای داده‌های آموزشی با k=30

	precision	recall	f1-score	support
1.0	0.52	1.00	0.68	163
2.0	0.00	0.00	0.00	32
3.0	0.00	0.00	0.00	10
4.0	0.00	0.00	0.00	11
5.0	0.00	0.00	0.00	12
6.0	0.00	0.00	0.00	18
7.0	0.00	0.00	0.00	3
9.0	0.00	0.00	0.00	9
10.0	1.00	0.03	0.05	36
14.0	0.00	0.00	0.00	3
15.0	0.00	0.00	0.00	4
16.0	0.00	0.00	0.00	15
accuracy			0.52	316
macro avg	0.13	0.09	0.06	316
weighted avg	0.38	0.52	0.36	316

شکل 3-9: Classification Report برای داده‌های تست با k=30

شکل 3-8 و 3-9 نشان می‌دهد که داده‌های آموزشی دارای دقت پایین‌تری از داده‌های تست هستند. وقتی این حالت رخ می‌دهد یعنی مدل underfit شده است.

****

ج) به کمک روش k-fold cross validation مقدار بهینه k و بهترین معیار فاصله (کسینوسی، اقلیدوسی و منهتن) را به دست آورید. حال این الگوریتم را به ازای این مقادیر بهینه بر روی دیتاست دانلود شده اعمال کنید. تمام معیارهای آموزش را برای داده‌های آموزش و تست به دست آورید و آن‌ها را تحلیل کنید.

در این قسمت مقدار بهینه و بهترین معیار فاصله (کسینوسی، اقلیدوسی و منهتن) را با استفاده از Gridsearch به دست می‌آوریم. هنگام استفاده از این روش تعداد همسایه‌ها را عددی بین 1 تا 30 در نظر می‌گیریم. چون

در بخش قبل دیدیم که با مقدار 1 مدل overfit شده و با مقدار 30 مدل underfit شده است. پس مقدار بهینه عددی بین این دو خواهد بود. با اجرای این الگوریتم بهترین پارامترها به صورت زیر است:

```
{'metric': 'cosine', 'n_neighbors': 10}
```

یعنی بهترین معیار فاصله معیار کسینوسی و بهترین شعاع همسایگی 10 است. حال مدل را با استفاده از این معیارها fit می‌کنیم و بعد بر روی داده‌های تست predict می‌کنیم.

Matrix و Classification Report برای داده‌های تست به صورت زیر است:

```
[[79  2  0  0  0  0  0  0  1  0  0  0]
 [10  2  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  5  0  0  0  0  0  0  0  0  0]
 [ 1  0  0  3  0  0  0  0  0  0  0  0]
 [ 1  0  0  0  0  0  0  0  0  0  0  0]
 [ 7  0  0  0  0  0  0  0  0  0  0  0]
 [ 2  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0]
 [ 9  0  0  0  0  0  0  0  5  0  0  0]
 [ 1  0  0  0  0  0  0  0  0  0  0  0]
 [ 1  0  0  0  0  0  0  0  0  0  0  0]
 [ 5  1  0  0  0  0  0  1  0  0  0  0]]
```

شکل 3-10: Confusion Matrix برای داده‌های آموزشی با k=10

	precision	recall	f1-score	support
1.0	0.68	0.96	0.80	82
2.0	0.40	0.17	0.24	12
3.0	1.00	1.00	1.00	5
4.0	1.00	0.75	0.86	4
5.0	0.00	0.00	0.00	1
6.0	0.00	0.00	0.00	7
8.0	0.00	0.00	0.00	2
9.0	0.00	0.00	0.00	0
10.0	0.83	0.36	0.50	14
14.0	0.00	0.00	0.00	1
15.0	0.00	0.00	0.00	1
16.0	0.00	0.00	0.00	7
accuracy			0.69	136
macro avg	0.33	0.27	0.28	136
weighted avg	0.60	0.69	0.62	136

شکل 3-11: Classification Report برای داده‌های تست با k=10

با توجه به اینکه داده‌ها بالانس نیستند بهترین نتیجه همین نتیجه‌ای است که در شکل 3-11 و 3-10 نشان داده شده است. در این شکل‌ها می‌بینیم که مقادیر معیارها برای اکثر کلاس‌ها دارای مقدار خوبی هستند و دقت نیز 69 درصد است. در این حالت مدل به خوبی آموزش دیده است.

منابع:

[1] Larose, Daniel T., and Chantal D. Larose. Discovering knowledge in data: an introduction to data mining. Vol. 4. John Wiley & Sons, 2014.

[2] Shmueli, Galit, Nitin R. Patel, and Peter C. Bruce. Data mining for business intelligence: Concepts, techniques, and applications in Microsoft Office Excel with XLMiner. John Wiley and Sons, 2011.

[3] Larose, Daniel T. Discovering Statistics. Macmillan Higher Education, 2011.