به نام خدا

پروژه داده‌کاوی

نام استاد : دکتر عادل محمد پور

گردآورنده : محمّد لشکری

شماره دانشجویی : ۹۵۱۳۴۳۵

بهار ۹۹

سؤال ۱

قسمت (آ):

در مورد اینکه کدام روش برای جایگزینی مقادیر گمشده مناسب تر است نمی توان نظر کلی داد، اگر انحراف معیار داده ها کم باشد جایگزینی با میانگین کار معقولی است. جایگزینی مقادیر گمشده با استفاده از KNN Imputer با مقدار k مناسب، کار معقول تری به نظر می رسد البته مقدار k مناسب برای همه دیتاست ها یکسان نیست و این مسئله همچنان باز است.

در اینجا نتایج حاصل از Simple Imputer و KNN Imputer را مشاهده می کنید؛ از اعمال روش های حذف رکورد یا حذف ویژگی بر روی دیتاست صرف نظر شده و نتایج حاصل از KNN روی دیتاست اعمال شده است.

KNN Imputer (n_neighbors = 5)	Simple Imputer
D14 = 0.324752 D17 = 1.528718 E7 = 2.168178e+07 F13 = 125919.6	D14 = 1.142708 D17 = 1.142708 E7 = 1.565725e+08 F13 = 1193427.5

	Country Name	Country Code	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	2.168178e+07	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	125919.6	C	3075
10	Morocco	MAR	0.324752	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	1.528718	9.903122e+06	407310.0	C	18640
14	India	IND	1.148215	1.324171e+09	2973190.0	B	26917

Main Table 1: Missing values handled

قسمت (ب):

به دلایل مختلفی Categorical variables را به داده‌های عددی تبدیل می‌کنیم. تعدادی از این دلایل در زیر ذکر شده‌اند:

۱ـ این متغیر ها را نمی‌توان برای محاسبات KNN استفاده نمود. (زیرا فاصله در نوعی از متغیرها معنی ندارند.)

۲ـ این متغیر ها را نمی‌توان برای رگریسیون لجستیک یا رگریسیون خطی استفاده نمود.

۳ـ داده‌های عددی مقدار حافظه کمتری را در Secondary storage, RAM به خود اختصاص می‌دهند.

در جدولی که مشاهده می‌کنید داده‌های موجود در ستون‌های **CountryCode** و **International Visitors** با استفاده از کلاس LabelEncoder به داده‌های عددی تبدیل شده‌اند.

	Country Name	Country Code	Population growth	Total population	Area (sq. km)	International Visitors	Coronavirus Cases
0	Brazil	0	0.817556	2.076529e+08	8358140.0	1	59324
1	Switzerland	1	1.077221	8.372098e+06	39516.0	1	29061
2	Germany	2	1.193867	8.266768e+07	348900.0	0	156727
3	Denmark	3	0.834638	2.168178e+07	42262.0	1	8575
4	Spain	4	-0.008048	4.644396e+07	500210.0	0	223759
5	France	5	0.407491	6.689611e+07	547557.0	0	161488
6	Japan	9	-0.115284	1.269945e+08	364560.0	1	13231
7	Greece	6	-0.687543	1.074674e+07	128900.0	2	2506
8	Iran	8	1.148789	8.027743e+07	1628760.0	3	90481
9	Kuwait	10	2.924206	4.052584e+06	125919.6	2	3075
10	Morocco	11	0.324752	3.527679e+07	446300.0	2	4047
11	Nigeria	12	2.619034	1.859896e+08	910770.0	3	1182
12	Qatar	13	3.495070	2.569804e+06	11610.0	1	10287
13	Sweden	14	1.528718	9.903122e+06	407310.0	2	18640
14	India	7	1.148215	1.324171e+09	2973190.0	1	26917

Main Table 2: Country Code and Inter. Visitors converted

قسمت (ج):

بله، به Feature scaling نیاز داریم زیرا محدوده داده‌ها در ستون‌های .Total pop و pop. Growth و Area وسیع می‌باشد و ممکن است داده های پرت معتبر وجود داشته باشد. در جداول زیر نتایج حاصل از MinMaxScaler و StandardScaler و Normalizer را مشاهده می‌کنید. نتایج حاصل از StandardScaler حاوی مقادیر منفی است در حالی که داده‌های اصلی همگی مثبت هستند. پس نتایج بدست آمده معتبر نیست از طرفی Normalizer فراوانی داده ها را تغییر می‌دهد(تابع چگالی عوض می‌شود) و در نتیجه مدلی که به دست می‌آید لزوماً معتبر نیست؛ پس برای این دیتاست به کارگیری MinMaxScaler معقول‌تر است.

	Population growth	Total population	Coronavirus Cases	Area (sq. km)
0	3.933941e-09	0.999191	0.000285	0.040218
1	1.286658e-07	0.999983	0.003471	0.004720
2	1.444160e-08	0.999989	0.001896	0.004220
3	3.849480e-08	0.999998	0.000395	0.001949
4	-1.732739e-10	0.999930	0.004817	0.010769
5	6.091179e-09	0.999964	0.002414	0.008185
6	-9.077849e-10	0.999996	0.000104	0.002871
7	-6.397224e-08	0.999928	0.000233	0.011993
8	1.430728e-08	0.999794	0.001127	0.020285
9	7.212176e-07	0.999517	0.000758	0.031056
10	9.205086e-09	0.999920	0.000115	0.012650
11	1.408144e-08	0.999988	0.000006	0.004897
12	1.360028e-06	0.999982	0.004003	0.004518
13	1.542367e-07	0.999153	0.001881	0.041095
14	8.671172e-10	0.999997	0.000020	0.002245

Normalizer

	Population growth	Total population	Coronavirus Cases	Area (sq. km)
0	0.359846	0.155178	0.261222	1.000000
1	0.421929	0.004390	0.125256	0.003343
2	0.449817	0.060607	0.698837	0.040411
3	0.363930	0.014461	0.033215	0.003672
4	0.162457	0.033198	1.000000	0.058539
5	0.261806	0.048673	0.720227	0.064212
6	0.136818	0.094147	0.054134	0.042287
7	0.000000	0.006187	0.005949	0.014053
8	0.439039	0.058798	0.401205	0.193751
9	0.863515	0.001122	0.008505	0.013695
10	0.242024	0.024748	0.012872	0.052080
11	0.790553	0.138786	0.000000	0.107729
12	1.000000	0.000000	0.040907	0.000000
13	0.529875	0.005549	0.078436	0.047409
14	0.438902	1.000000	0.115623	0.354828

MinMaxScaler

	Population growth	Total population	Coronavirus Cases	Area (sq. km)
0	-0.265443	0.187332	0.078032	3.489602
1	-0.032864	-0.434106	-0.361669	-0.522168
2	0.071614	-0.202422	1.493232	-0.372964
3	-0.250143	-0.392601	-0.659316	-0.520844
4	-1.004926	-0.315382	2.467161	-0.299992
5	-0.632733	-0.251604	1.562406	-0.277159
6	-1.100976	-0.064193	-0.591668	-0.365411
7	-1.613541	-0.426701	-0.747495	-0.479062
8	0.031239	-0.209876	0.530722	0.244266
9	1.621459	-0.447576	-0.739228	-0.480499
10	-0.706841	-0.350206	-0.725105	-0.325991
11	1.348119	0.119778	-0.766732	-0.101994
12	2.132775	-0.452200	-0.634442	-0.535626
13	0.371537	-0.429331	-0.513079	-0.344795
14	0.030724	3.669089	-0.392820	0.892636

StandardScaler

	Country Name	Country Code	Population growth	Total population	Area (sq. km)	International Visitors	Coronavirus Cases
0	Brazil	0	0.359846	0.155178	1.000000	1	0.261222
1	Switzerland	1	0.421929	0.004390	0.003343	1	0.125256
2	Germany	2	0.449817	0.060607	0.040411	0	0.698837
3	Denmark	3	0.363930	0.014461	0.003672	1	0.033215
4	Spain	4	0.162457	0.033198	0.058539	0	1.000000
5	France	5	0.261806	0.048673	0.064212	0	0.720227
6	Japan	9	0.136818	0.094147	0.042287	1	0.054134
7	Greece	6	0.000000	0.006187	0.014053	2	0.005949
8	Iran	8	0.439039	0.058798	0.193751	3	0.401205
9	Kuwait	10	0.863515	0.001122	0.013695	2	0.008505
10	Morocco	11	0.242024	0.024748	0.052080	2	0.012872
11	Nigeria	12	0.790553	0.138786	0.107729	3	0.000000
12	Qatar	13	1.000000	0.000000	0.000000	1	0.040907
13	Sweden	14	0.529875	0.005549	0.047409	2	0.078436
14	India	7	0.438902	1.000000	0.354828	1	0.115623

Main Table 3: Min Max Scaler applied to the dataset

قسمت (د):

در جدول Outliers داده های پرت را مشاهده می‌کنید. این داده ها از دیتاست حذف شده اند اما ممکن است داده های پرت معتبر وجود داشته باشد و به همین دلیل حذف آنها کار درستی نیست و باید از روش های نرمال سازی استفاده کنیم تا میزان تأثیر متغیر ها روی نتیجه استاندارد شود که یکی از این روش ها در قسمت (ب) بر روی دیتاست اعمال شده است. (ستون هایی که در جدول Outliers نیستند حاوی داده پرت نبوده اند.)

	Population growth	Total population	Area (sq. km)
0	NaN	NaN	1.000000
1	NaN	NaN	NaN
2	NaN	NaN	NaN
3	NaN	NaN	NaN
4	NaN	NaN	NaN
5	NaN	NaN	NaN
6	NaN	NaN	NaN
7	NaN	NaN	NaN
8	NaN	NaN	NaN
9	0.863515	NaN	NaN
10	NaN	NaN	NaN
11	NaN	NaN	NaN
12	1.000000	NaN	NaN
13	NaN	NaN	NaN
14	NaN	1.000000	0.354828

Qutliers

	Country Name	Country Code	Population growth	Total population	Area (sq. km)	International Visitors	Coronavirus Cases
1	Switzerland	1	0.421929	0.004390	0.003343	1	0.125256
2	Germany	2	0.449817	0.060607	0.040411	0	0.698837
3	Denmark	3	0.363930	0.014461	0.003672	1	0.033215
4	Spain	4	0.162457	0.033198	0.058539	0	1.000000
5	France	5	0.261806	0.048673	0.064212	0	0.720227
6	Japan	9	0.136818	0.094147	0.042287	1	0.054134
7	Greece	6	0.000000	0.006187	0.014053	2	0.005949
8	Iran	8	0.439039	0.058798	0.193751	3	0.401205
10	Morocco	11	0.242024	0.024748	0.052080	2	0.012872
11	Nigeria	12	0.790553	0.138786	0.107729	3	0.000000
13	Sweden	14	0.529875	0.005549	0.047409	2	0.078436

Main Table 4: Outliers removed

قسمت (ه):

نتایجی که مشاهده می‌کنید حاصل از رگرسیون خطی چند گانه بر روی Main Table 4 است که در آن **coronavirus cases** متغیر وابسته و **Population growth** و **Total population** و **Area (sq. km)** متغیر های مستقل هستند. همان طور که مشاهده می‌شود P- مقدار ها از ۰٫۰۵ بیشتر هستند (با فرض alpha=0.05) پس این خط، خط خوبی نیست و همچنین ضریب همبستگی نیز به صفر نزدیک است که دلالت بر مناسب نبودن مدل دارد.

Dep. Variable:	Q("Coronavirus Cases")	**R-squared:**	0.104
Model:	OLS	**Adj. R-squared:**	0.280
Method:	Least Squares	**F-statistic:**	0.2700
Date:	Thu, 14 May 2020	**Prob (F-statistic):**	0.845
Time:	16:34:27	**Log-Likelihood:**	-3.2752
No. Observations:	11	**AIC:**	14.55
Df Residuals:	7	**BIC:**	16.14
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	0.3264	0.249	1.311	0.231	-0.262	0.915
Q("Population growth")	-0.4321	0.677	-0.638	0.544	-2.033	1.168
Q("Total population")	-0.2758	3.741	-0.074	0.943	-9.122	8.570
Q("Area (sq. km)")	2.0968	2.812	0.746	0.480	-4.553	8.746

Omnibus:	2.400	**Durbin-Watson:**	1.612
Prob(Omnibus):	0.301	**Jarque-Bera (JB):**	1.643
Skew:	0.859	**Prob(JB):**	0.440
Kurtosis:	2.203	**Cond. No.**	34.8

MSE	10032800065.171135
RMSE	100163.866065

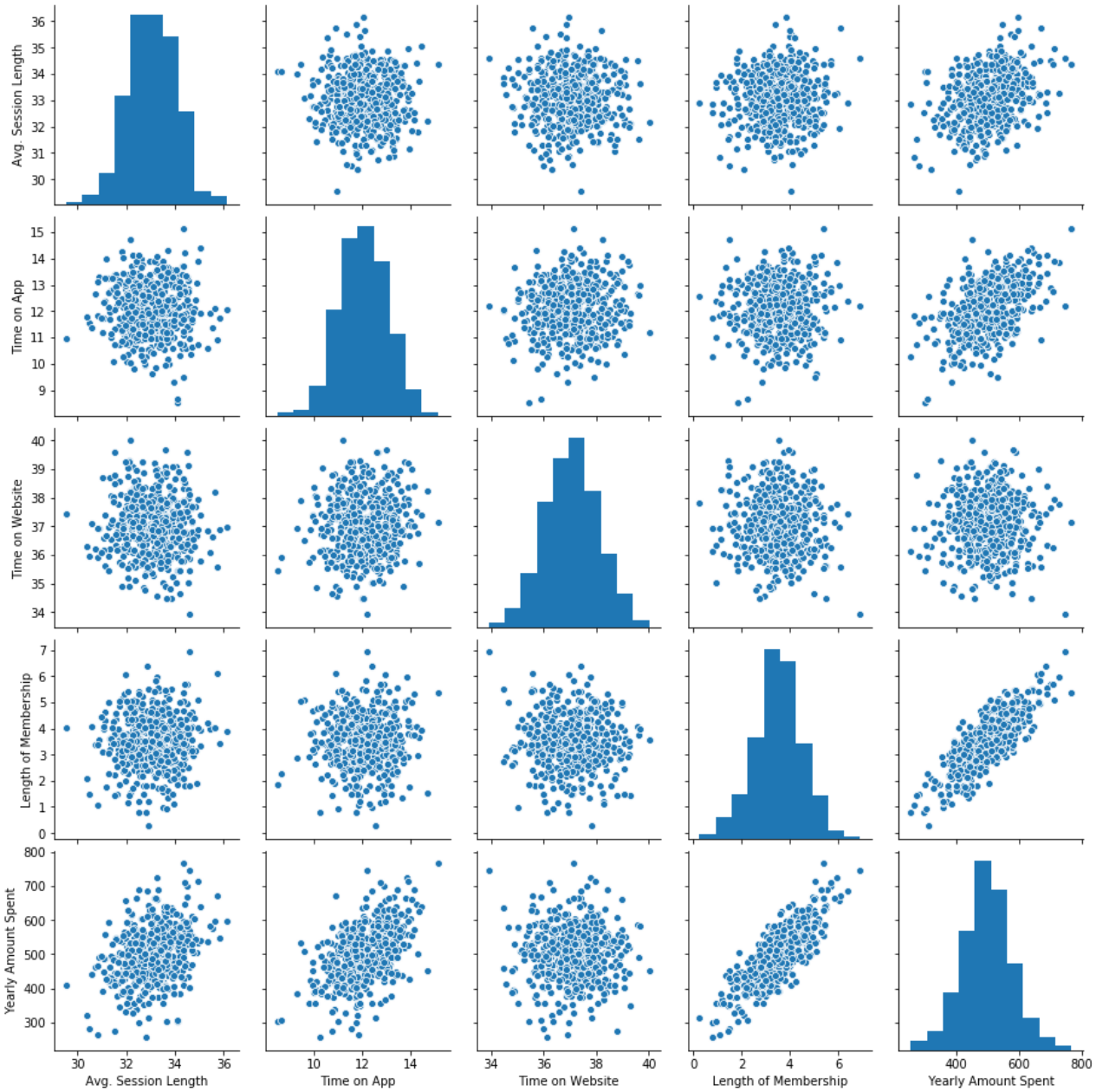
قسمت (و):

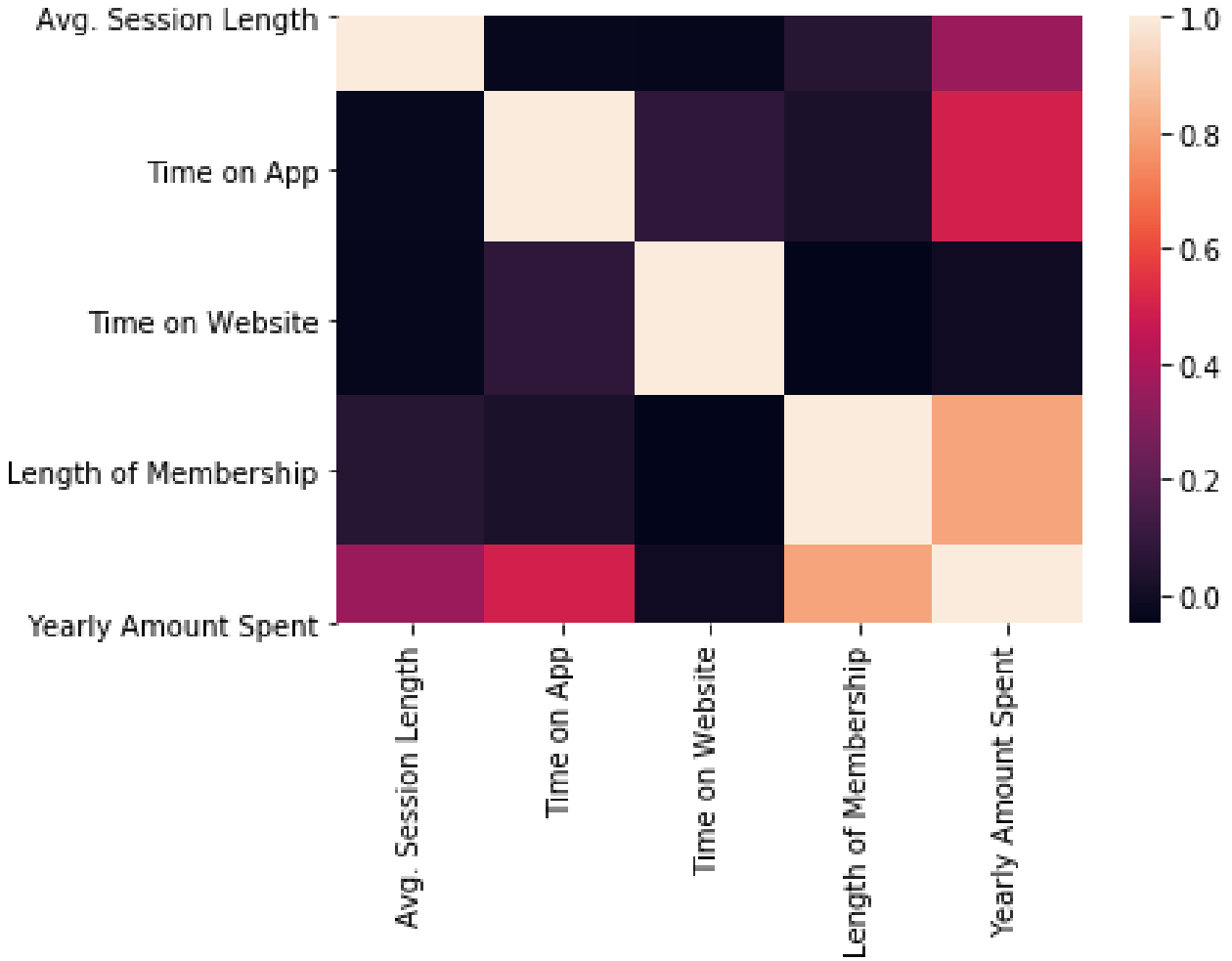
	Parameter value
a (X^2 coef)	-6.103689
b (X coef)	4.264761
c (intercept)	-0.041403

سؤال ۲

قسمت (آ):

ویژگی های **Yearly Amount of Spent** و **Length of Membership** به هم هبستگی بیشتری نسبت به سایر ویژگی ها دارند.




Avg. Session Length
Time on App
Time on Website
Length of Membership
Yearly Amount Spent
Avg. Session Length
Time on App
Time on Website
Length of Membership
Yearly Amount Spent
1.0
0.8
0.6
0.4
0.2
0.0

قسمت (ب):

نتایجی که مشاهده می‌کنید حاصل از رگرسیون خطی چندگانه بر روی دیتاست مربوطه است که در آن **Yearly Amount of Spent** متغیر وابسته و **Time on App** و **Time on Website** و **Length of Membership** متغیر های مستقل هستند؛ فقط P-مقدار مربوط به **Time on Website** از مقدار آلفا (با فرض alpha=0.05) بیشتر است که فرض H0 رد نمی‌شود.

Dep. Variable:	Q("Yearly Amount Spent")	**R-squared:**	0.881
Model:	OLS	**Adj. R-squared:**	0.880
Method:	Least Squares	**F-statistic:**	1226.
Date:	Sun, 17 May 2020	**Prob (F-statistic):**	6.48e-229
Time:	17:40:19	**Log-Likelihood:**	-2363.2
No. Observations:	500	**AIC:**	4734.
Df Residuals:	496	**BIC:**	4751.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-169.9607	46.820	-3.630	0.000	-261.950	-77.971
Q("Time on App")	38.0130	1.240	30.662	0.000	35.577	40.449
Q("Time on Website")	-0.3195	1.221	-0.262	0.794	-2.718	2.079
Q("Length of Membership")	63.1011	1.231	51.274	0.000	60.683	65.519

Omnibus:	1.264	**Durbin-Watson:**	2.075
Prob(Omnibus):	0.532	**Jarque-Bera (JB):**	1.358
Skew:	-0.108	**Prob(JB):**	0.507
Kurtosis:	2.863	**Cond. No.**	1.50e+03

MSE	746.094151
RMSE	27.314724

قسمت (ج):

در زیر تقریب خطای تست با 5-fold cross validation را مشاهده می‌کنید:

5-fold cross validation
Test set MSE ≅ 758.5485675790217

قسمت (د):

خطای داده های تست و آموزش نزدیک به هم هستند (مقدار هر دو کوچک است) و underfit رخ داده است. bias این مدل کم (نزدیک به صفر) و variance آن زیاد است. با توجه به این که RMSE داده های تست و آموزش کم است مدل خوب است.

Test set RMSE (estimated)	27.541760429918448
Train set RMSE	27.314724070910714

قسمت (ه):

با توجه به کارکرد مدل که در قسمت (ج) بررسی شد و ضرایب به دست آمده در قسمت (ب) زمان استفاده از اپلیکیشن موبایل بیشتر بوده و با **Yearly Amount of Spent** نسبت مستقیم دارد، پس پیشنهاد می‌شود روی اپلیکیشن تمرکز کنند.

سؤال ۳

قسمت (آ):

داده های گمشده با نتایج حاصل از KNN_Imputer (n_neighbors = 5) جایگزین شده اند. در زیر ۱۰ ویژگی ۵ داده اول دیتاست را بعد از جایگزینی مقادیر گمشده مشاهده می‌کنید.

	0	1	2	3	4	5	6	7	8	9	10
0	75.0	0.0	190.0	80.0	91.0	193.0	371.0	174.0	121.0	-16.0	13.0
1	56.0	1.0	165.0	64.0	81.0	174.0	401.0	149.0	39.0	25.0	37.0
2	54.0	0.0	172.0	95.0	138.0	163.0	386.0	185.0	102.0	96.0	34.0
3	55.0	0.0	175.0	94.0	100.0	202.0	380.0	179.0	143.0	28.0	11.0
4	75.0	0.0	190.0	80.0	88.0	181.0	360.0	177.0	103.0	-16.0	13.0

قسمت (ب):

دیتاست قبل از انجام این قسمت با استفاده از Standard Scaler مقیاس شده است. در دو جدول زیر تعداد مقادیر موجود از هر کلاس را برای داده های آموزش و تست مشاهده می‌کنید(با نظر گرفتن ۲۰ درصد از داده های اصلی به عنوان تست و ۸۰ درصد به عنوان داده های آموزش)

	Test set	Train set
1	46	199
2	12	32
3	4	11
4	2	13
5	2	11
6	7	18
7	1	2
8	0	2
9	1	8
10	10	40
11	0	0
12	0	0
13	0	0
14	1	3
15	0	5
16	5	17

	1	2	3	4	5	6	7	8	9	10	14	16
1	40	9	2	1	1	4	1	1	3	1	0	5
2	1	2	0	0	0	0	0	0	0	0	0	1
3	0	0	2	0	0	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	1	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0
6	1	0	0	0	0	1	0	0	0	0	2	0
7	2	0	0	0	0	2	0	0	1	0	0	2
8	0	0	0	0	0	0	0	0	0	0	0	0
9	1	0	0	0	0	0	0	0	5	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0
14	0	1	0	0	0	0	0	0	0	0	0	0
16	2	0	0	0	1	1	0	0	0	0	0	0

Confusion matrix for test set: k = 1

	precision	recall	f1-score	support
1	0.87	0.62	0.72	65
2	0.17	0.50	0.25	4
3	0.50	1.00	0.67	2
4	0.50	0.50	0.50	2
5	0.00	0.00	0.00	0
6	0.29	0.29	0.29	7
7	0.00	0.00	0.00	0
9	0.00	0.00	0.00	0
10	0.50	0.83	0.62	6
14	0.00	0.00	0.00	0
15	0.00	0.00	0.00	1
16	0.00	0.00	0.00	4
accuracy			0.57	91
macro avg	0.24	0.31	0.25	91
weighted avg	0.71	0.57	0.61	91

Classification report for test set: k = 1

	1	2	3	4	5	6	7	9	10	14	16
1	46	12	4	2	2	7	1	1	10	1	5
2	0	0	0	0	0	0	0	0	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
5	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0

Confusion matrix for test set: k = 30

	precision	recall	f1-score	support
1	1.00	0.51	0.67	91
2	0.00	0.00	0.00	0
3	0.00	0.00	0.00	0
4	0.00	0.00	0.00	0
5	0.00	0.00	0.00	0
6	0.00	0.00	0.00	0
7	0.00	0.00	0.00	0
9	0.00	0.00	0.00	0
10	0.00	0.00	0.00	0
14	0.00	0.00	0.00	0
15	0.00	0.00	0.00	0
16	0.00	0.00	0.00	0
accuracy			0.51	91
macro avg	0.09	0.05	0.06	91
weighted avg	1.00	0.51	0.67	91

Classification report for test set: k = 30

	1	2	3	4	5	6	7	8	9	10	14	15	16
1	199	0	0	0	0	0	0	0	0	0	0	0	0
2	0	32	0	0	0	0	0	0	0	0	0	0	0
3	0	0	11	0	0	0	0	0	0	0	0	0	0
4	0	0	0	13	0	0	0	0	0	0	0	0	0
5	0	0	0	0	11	0	0	0	0	0	0	0	0
6	0	0	0	0	0	18	0	0	0	0	0	0	0
7	0	0	0	0	0	0	2	0	0	0	0	0	0
8	0	0	0	0	0	0	0	2	0	0	0	0	0
9	0	0	0	0	0	0	0	0	8	0	0	0	0
10	0	0	0	0	0	0	0	0	0	40	0	0	0
14	0	0	0	0	0	0	0	0	0	0	3	0	0
15	0	0	0	0	0	0	0	0	0	0	0	5	0
16	0	0	0	0	0	0	0	0	0	0	0	0	17

Confusion matrix for train set: k = 1

	precision	recall	f1-score	support
1	1.00	1.00	1.00	199
2	1.00	1.00	1.00	32
3	1.00	1.00	1.00	11
4	1.00	1.00	1.00	13
5	1.00	1.00	1.00	11
6	1.00	1.00	1.00	18
7	1.00	1.00	1.00	2
8	1.00	1.00	1.00	2
9	1.00	1.00	1.00	8
10	1.00	1.00	1.00	40
14	1.00	1.00	1.00	3
15	1.00	1.00	1.00	5
16	1.00	1.00	1.00	17
accuracy			1.00	361
macro avg	1.00	1.00	1.00	361
weighted avg	1.00	1.00	1.00	361

Classification report for train set: k = 1

	1	2	3	4	5	6	7	8	9	10	14	15	16
1	199	32	11	13	11	18	2	2	8	39	3	5	17
2	0	0	0	0	0	0	0	0	0	0	0	0	0
3	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
5	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
7	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
9	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	1	0	0	0
14	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
16	0	0	0	0	0	0	0	0	0	0	0	0	0

Confusion matrix for train set: k = 30

	precision	recall	f1-score	support
1	1.00	0.55	0.71	360
2	0.00	0.00	0.00	0
3	0.00	0.00	0.00	0
4	0.00	0.00	0.00	0
5	0.00	0.00	0.00	0
6	0.00	0.00	0.00	0
7	0.00	0.00	0.00	0
8	0.00	0.00	0.00	0
9	0.00	0.00	0.00	0
10	0.03	1.00	0.05	1
14	0.00	0.00	0.00	0
15	0.00	0.00	0.00	0
16	0.00	0.00	0.00	0
accuracy			0.55	361
macro avg	0.08	0.12	0.06	361
weighted avg	1.00	0.55	0.71	361

Classification report for train set: k = 30

تحلیل نتایج قبل:

k = 1

برای داده های آموزش کاملا درست پیش بینی انجام شده است اما برای داده های تست recall برای کلاس ۱خوب است. مقدار precision برای کلاس های ۳و۴و۱۰ در داده های تست از کلاس های دیگر بیشتر بوده اما خوب نیست.(برای کلاس های ۲ تا ۱۵ مقدار precision مهم است.)

k = 30

همان طور که در نتایج قبل مشاهده می‌شود برای داده های تست، کلاس همه داده ها ۱ پیش بینی شده است و چون کلاس ۱ بیانگر نرمال بودن است، اینکه یک فرد به اشتباه نرمال تشخیص داده شود خسارت زیادی وارد می‌کند. مقدار recall برای کلاس ۱ به ۰٫۵ نزدیک است و برای کلاس های دیگر مقدار precision صفر است پس این مدل، مدل خوبی نیست و از مدل قبل برای داده های تست بدتر عمل کرده است.
برای داده های آموزش نیز با توجه به مقادیر recall و precision و f1-score از مدل قبل بدتر عمل کرده است.

قسمت (ج):

بهترین مدل:

```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='manhattan'
, metric_params=None, n_jobs=None, n_neighbors=3, p=2,
                     weights='uniform')
```

	1	2	3	4	5	6	7	8	9	10	14	15
1	46	10	3	1	1	7	1	1	10	1	0	4
2	0	2	0	0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0
7	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
9	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	1
15	0	0	0	0	0	0	0	0	0	0	0	0

Confusion matrix for test set: best estimator

	precision	recall	f1-score	support
1	1.00	0.54	0.70	85
2	0.17	1.00	0.29	2
3	0.25	1.00	0.40	1
4	0.50	1.00	0.67	1
5	0.50	1.00	0.67	1
6	0.00	0.00	0.00	0
7	0.00	0.00	0.00	0
9	0.00	0.00	0.00	0
10	0.00	0.00	0.00	0
14	0.00	0.00	0.00	0
15	0.00	0.00	0.00	1
16	0.00	0.00	0.00	0
accuracy			0.56	91
macro avg	0.20	0.38	0.23	91
weighted avg	0.95	0.56	0.68	91

Classification report for test set: best estimator

تحلیل نتیجه:

مقدار recall برای کلاس ۱ همانند مدل های قبل خوب نیست و مقدار recall برای کلاس های ۲و۳و۴و۵ خوب است اما برای کلاس های ۲ تا ۱۵ مقدار precision مهم است؛ مقدار precision خوب نیست اما از مدل های قبل بهتر است.