سوگند رضوی – امین عرفانیان

۹۹۱۲۷۶۲۷۵۶- ۹۷۲۲۷۶۲۱۰۰

لینک پروژه:

https://colab.research.google.com/drive/1f_dRNrlHHY96ImBqw6YYSU2pJrP0kb_1?usp=sharing

فاز ۱

سوال ۳:

برای پیش بینی حمله قلبی باید correlation بین آن feature و مابقی را حساب کرد.

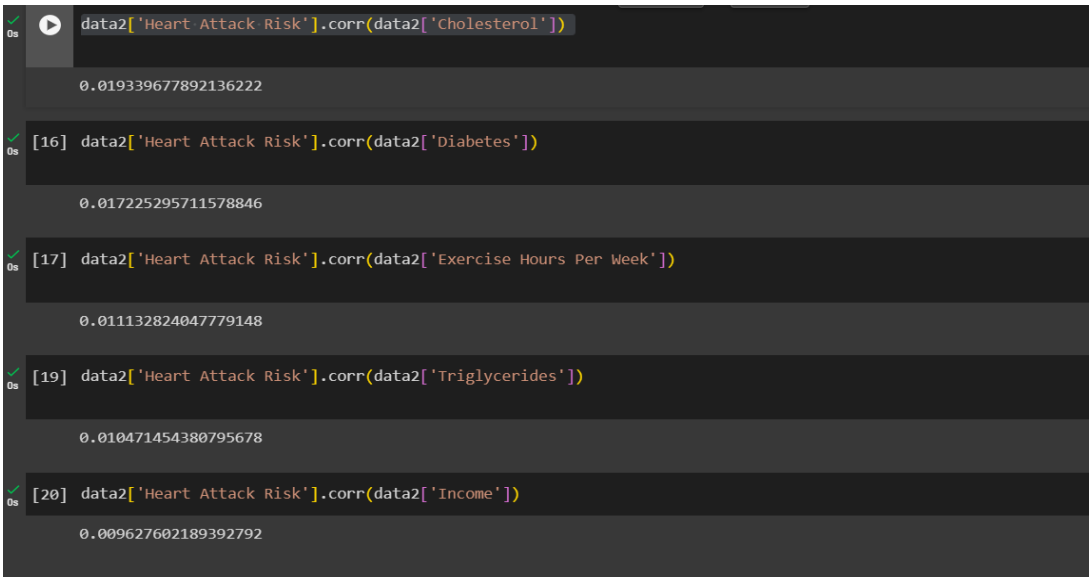
```
# CORRELATION
harcorr = data2.corr().sort_values(by='Heart Attack Risk',ascending=False)
columns_corr = harcorr.index[1:7]
columns_corr
```

خروجی:

```
Index(['Cholesterol', 'Diabetes', 'Exercise Hours Per Week', 'Triglycerides',
       'Income', 'Age'],
      dtype='object')
```

متوجه میشیم که مثلا cholesterol ریسک حمله قلبی را بالا میبرد و به همین ترتیب دیابت و غیره

حالا برای correlation بین خروجی های بالا و heart attack risk داریم:



پس نتیجه میگیریم از بین همه feature هایی که در دیتاست داشتیم این موارد ریسک حمله قلبی را افزایش میدهند.

"ما صرفا برای ۷ مورد اولیه بررسی کردیم. که هرچه بیشتر میشد correlation هم کمتر میشد و ریسک را کاهش میداد. یعنی از همه بیشتر کلسترول و در اینجا از همه کمتر درآمد(income) هست."

فاز ۲

سوال ۱:

به جز accuracy و currentness که نمیتوان به دست آورد مابقی داخل کد هست و برای consistency داریم:

Feature name	consistency
Patient id	1
Age	1
Sex	1

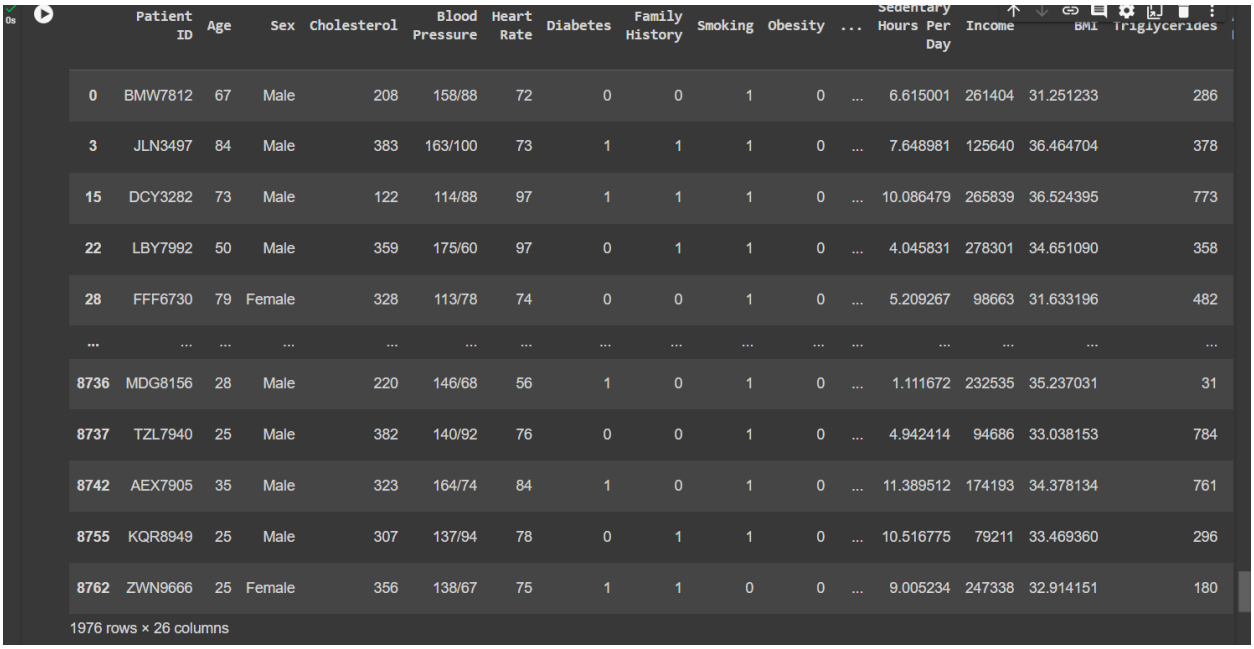
Cholesterol	0 (with blood pressure,num6)
Blood Pressure	0 (with cholesterol, num6 , num3)
Heart Rate	1
Diabetes	1
Family history	1
smoking	1
obesity	0(with BMI, num1)
Alcohol Consumption	1
Exercise Hours Per Week	0(with physical activity days per week, num2)
Diet	1
Previous Heart Problems	0(with medication use, Triglycerides, num4, num5)
Medication Use	0(with previous heart problems, num4)
Stress Level	1
Sedentary Hours Per Day	1
Income	1
BMI	0(with obesity, num1)
Triglycerides	0(with Previous Heart Problems, num5)
Physical Activity Days Per Week	0(with Exercise Hours Per Week, num2)
Sleep Hours Per Day	1

راهنمای جدول: در جدول آنهایی که سازگار هستند با عدد ۱ نمایش داده شده اند و آنهایی که ناسازگار هستند با عدد ۰ نمایش داده شدند و نوشته شده با چه چیزی ناسازگاری دارند و عدد مقابل آنها عددهای پایین هست که هم کد آن قرار داده شده هم تعداد ناسازگاری هایی که بین این دو وجود داشته.

1. BMI and Obesity

```
bmi_obes = data2[(data2['BMI'] > 30) & (data2['Obesity'] == 0)]
bmi_obes
```

output:



	Patient ID	Age	Sex	Cholesterol	Blood Pressure	Heart Rate	Diabetes	Family History	Smoking	Obesity	...	Sedentary Hours Per Day	Income	BMI	Triglycerides
0	BMW7812	67	Male	208	158/88	72	0	0	1	0	...	6.615001	261404	31.251233	286
3	JLN3497	84	Male	383	163/100	73	1	1	1	0	...	7.648981	125640	36.464704	378
15	DCY3282	73	Male	122	114/88	97	1	1	1	0	...	10.086479	265839	36.524395	773
22	LBY7992	50	Male	359	175/60	97	0	1	1	0	...	4.045831	278301	34.651090	358
28	FFF6730	79	Female	328	113/78	74	0	0	1	0	...	5.209267	98663	31.633196	482
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8736	MDG8156	28	Male	220	146/68	56	1	0	1	0	...	1.111672	232535	35.237031	31
8737	TZL7940	25	Male	382	140/92	76	0	0	1	0	...	4.942414	94686	33.038153	784
8742	AEX7905	35	Male	323	164/74	84	1	0	1	0	...	11.389512	174193	34.378134	761
8755	KQR8949	25	Male	307	137/94	78	0	1	1	0	...	10.516775	79211	33.469360	296
8762	ZWN9666	25	Female	356	138/67	75	1	1	0	0	...	9.005234	247338	32.914151	180

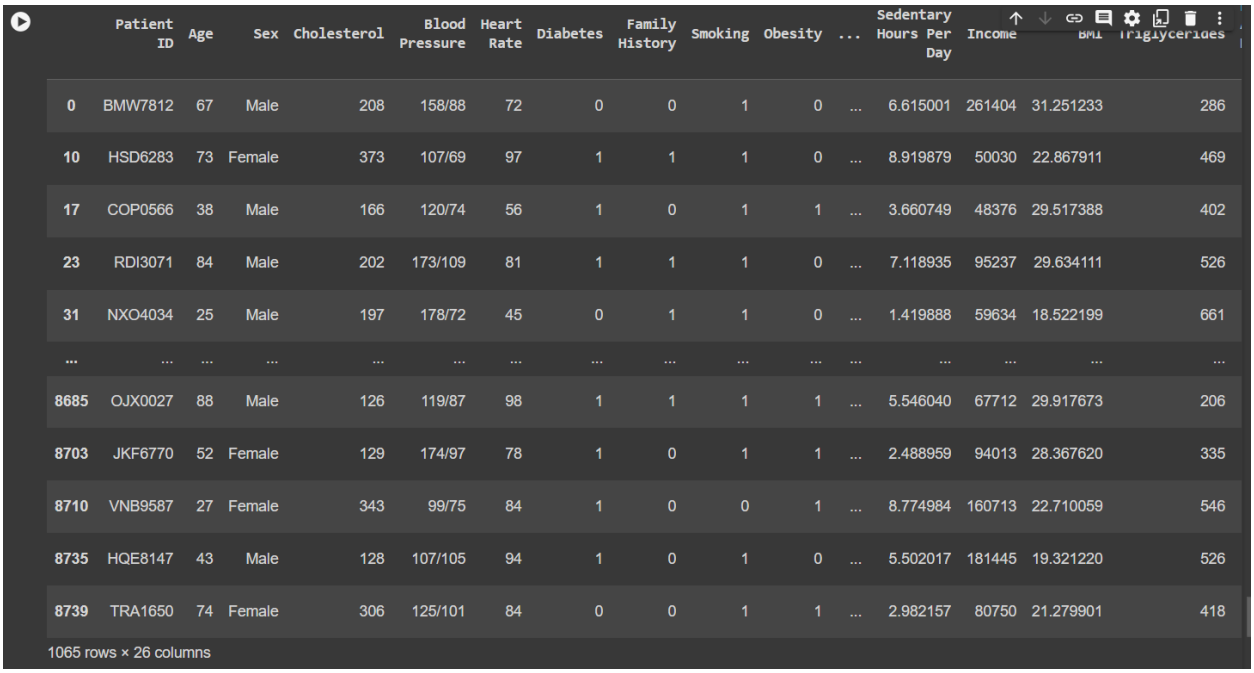
1976 rows × 26 columns

یعنی ۱۹۷۶ نفر در این مورد باهم ناسازگاری داشتند.

2. Exercise hours per week and physical activity days per week

```
pad_ehpw = data2[((data2['Physical Activity Days Per Week'] == 0) &
(data2['Exercise Hours Per Week'] > 0)) |
                 ((data2['Physical Activity Days Per Week'] == 1) &
(data2['Exercise Hours Per Week'] == 0))]
```

Output:



	Patient ID	Age	Sex	Cholesterol	Blood Pressure	Heart Rate	Diabetes	Family History	Smoking	Obesity	...	Sedentary Hours Per Day	Income	BMI	Triglycerides
0	BMW7812	67	Male	208	158/88	72	0	0	1	0	...	6.615001	261404	31.251233	286
10	HSD6283	73	Female	373	107/69	97	1	1	1	0	...	8.919879	50030	22.867911	469
17	COP0566	38	Male	166	120/74	56	1	0	1	1	...	3.660749	48376	29.517388	402
23	RDI3071	84	Male	202	173/109	81	1	1	1	0	...	7.118935	95237	29.634111	526
31	NXO4034	25	Male	197	178/72	45	0	1	1	0	...	1.419888	59634	18.522199	661
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8685	OJX0027	88	Male	126	119/87	98	1	1	1	1	...	5.546040	67712	29.917673	206
8703	JKF6770	52	Female	129	174/97	78	1	0	1	1	...	2.488959	94013	28.367620	335
8710	VNB9587	27	Female	343	99/75	84	1	0	0	1	...	8.774984	160713	22.710059	546
8735	HQE8147	43	Male	128	107/105	94	1	0	1	0	...	5.502017	181445	19.321220	526
8739	TRA1650	74	Female	306	125/101	84	0	0	1	1	...	2.982157	80750	21.279901	418

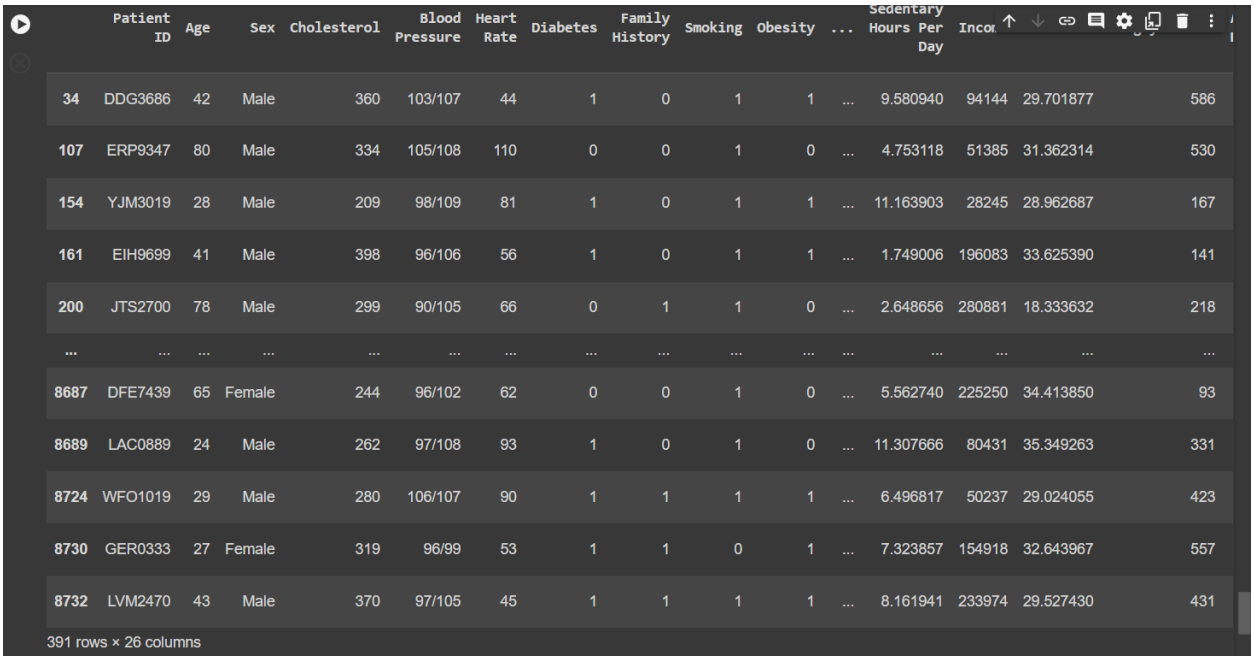
1065 rows × 26 columns

یعنی ۱۰۶۵ نفر باهم ناسازگاری دارند.

3.Blood pressure

```
bloodp = data2[data2['Blood Pressure'].str.split('/',
expand=True)[0].astype(int) < data2['Blood Pressure'].str.split('/',
expand=True)[1].astype(int)]
```

output:



	Patient ID	Age	Sex	Cholesterol	Blood Pressure	Heart Rate	Diabetes	Family History	Smoking	Obesity	...	Sedentary Hours Per Day	Incor		
34	DDG3686	42	Male	360	103/107	44	1	0	1	1	...	9.580940	94144	29.701877	586
107	ERP9347	80	Male	334	105/108	110	0	0	1	0	...	4.753118	51385	31.362314	530
154	YJM3019	28	Male	209	98/109	81	1	0	1	1	...	11.163903	28245	28.962687	167
161	EIH9699	41	Male	398	96/106	56	1	0	1	1	...	1.749006	196083	33.625390	141
200	JTS2700	78	Male	299	90/105	66	0	1	1	0	...	2.648656	280881	18.333632	218
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8687	DFE7439	65	Female	244	96/102	62	0	0	1	0	...	5.562740	225250	34.413850	93
8689	LAC0889	24	Male	262	97/108	93	1	0	1	0	...	11.307666	80431	35.349263	331
8724	WFO1019	29	Male	280	106/107	90	1	1	1	1	...	6.496817	50237	29.024055	423
8730	GER0333	27	Female	319	96/99	53	1	1	0	1	...	7.323857	154918	32.643967	557
8732	LVM2470	43	Male	370	97/105	45	1	1	1	1	...	8.161941	233974	29.527430	431

391 rows × 26 columns

یعنی ۳۹۱ نفر باهم ناسازگاری دارند.

4.Previous heart problems and medication use

```
php_Mu = data2[(data2['Previous Heart Problems'] == 1) &
(data2['Medication Use'] == 0) |
                          (data2['Previous Heart Problems'] == 0) &
(data2['Medication Use'] == 1)]
```

```
php_Mu
```

output:

	Patient ID	Age	Sex	Cholesterol	Blood Pressure	Heart Rate	Diabetes	Family History	Smoking	Obesity	...	Sedentary Hours Per Day	Income		
1	CZE1114	21	Male	389	165/93	98	1	1	1	1	...	4.963459	285768	27.194973	235
3	JLN3497	84	Male	383	163/100	73	1	1	1	0	...	7.648981	125640	36.464704	378
4	GFO8847	66	Male	318	91/88	93	1	1	1	1	...	1.514821	160555	21.809144	231
7	XXM0972	84	Male	220	131/68	107	0	0	1	1	...	10.543780	122093	22.221862	370
14	VTW9069	88	Male	297	112/81	102	1	1	1	0	...	10.425490	165300	25.491741	635
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8748	GQZ5013	81	Male	137	143/64	61	1	0	1	1	...	6.766948	50533	35.074391	191
8749	GNE7873	60	Male	156	111/96	83	1	1	1	0	...	6.189673	291495	26.605383	572
8752	UBM5982	65	Male	150	152/99	106	1	0	1	0	...	7.325356	55934	22.710546	83
8753	NVC8704	82	Male	311	126/108	87	0	1	1	1	...	8.402977	141521	27.694240	515
8760	XKA5925	47	Male	250	161/75	105	0	1	1	1	...	2.375214	36998	35.406146	527

4358 rows × 26 columns

یعنی ۴۳۵۸ نفر باهم ناسازگاری دارند.

5.Triglycerides and Previous Heart Problems

```
Triglycerides_pha = data2[(data2['Triglycerides'] < 150) &
(data2['Previous Heart Problems'] == 1)]
Triglycerides_pha
```

Output:

	Patient ID	Age	Sex	Cholesterol	Blood Pressure	Heart Rate	Diabetes	Family History	Smoking	Obesity	...	Sedentary Hours Per Day	Income	BMI	Triglycerides
16	DXB2434	69	Male	379	173/75	40	1	1	1	1	...	9.060509	267997	28.332747	68
53	IKY4481	67	Male	222	159/79	105	1	1	1	1	...	0.861540	286299	37.258748	92
59	WAR7163	72	Male	377	144/98	61	1	1	1	1	...	3.476703	249614	28.514638	106
64	TQT8266	53	Male	133	161/108	110	1	1	1	1	...	2.094265	182477	27.681792	67
75	DHP4080	55	Male	163	139/107	63	0	0	1	1	...	9.351067	158030	26.608767	131
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8741	NTL8842	45	Female	166	98/60	63	0	0	1	1	...	10.345259	87050	29.640828	74
8747	BBS4075	24	Male	396	118/89	45	0	0	1	1	...	11.467868	219922	38.436301	51
8752	UBM5982	65	Male	150	152/99	106	1	0	1	0	...	7.325356	55934	22.710546	83
8758	MSV9918	60	Male	121	94/76	61	1	1	1	0	...	10.806373	235420	19.655895	67
8761	EPE6801	36	Male	178	119/67	60	1	0	1	0	...	0.029104	209943	27.294020	114

702 rows × 26 columns

یعنی 702 نفر باهم ناسازگاری دارند.

6. Blood pressure and Cholesterol

```
bloodp = data2['Blood Pressure']
split_blood_pressure = bloodp.str.split('/', expand=True)
systolic_values = pd.to_numeric(split_blood_pressure[0])
diastolic_values = pd.to_numeric(split_blood_pressure[1])
Cholesterol_filtered_df = data2[(data2['Cholesterol'] > 240) &
(systolic_values < 120)  & (diastolic_values < 80)]
Cholesterol_filtered_df
```

Output:

	Patient ID	Age	Sex	Cholesterol	Blood Pressure	Heart Rate	Diabetes	Family History	Smoking	Obesity	...	Sedentary Hours Per Day	Income	BMI	Triglycerides
6	WYV0966	90	Male	358	102/73	84	0	0	1	0	...	0.627356	190450	28.885811	284
10	HSD6283	73	Female	373	107/69	97	1	1	1	0	...	8.919879	50030	22.867911	469
28	FFF6730	79	Female	328	113/78	74	0	0	1	0	...	5.209267	98663	31.633196	482
35	FLG2019	52	Female	360	94/60	106	1	0	1	1	...	7.695640	135099	27.095853	743
47	UBJ2564	70	Female	279	102/76	86	0	0	1	1	...	1.090400	191558	29.970809	792
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8705	GOR6731	76	Male	347	99/64	75	1	0	1	0	...	3.555172	102881	34.193579	696
8710	VNB9587	27	Female	343	99/75	84	1	0	0	1	...	8.774984	160713	22.710059	546
8722	PKV6924	26	Male	259	106/64	107	0	0	1	0	...	7.581411	76190	23.252114	653
8744	VXA0409	83	Male	322	91/69	67	1	1	1	0	...	11.155505	277472	28.873162	518
8757	YDX2478	59	Female	378	93/78	99	0	1	1	1	...	7.495231	70415	39.976061	158

661 rows × 26 columns

یعنی 661 نفر باهم ناسازگاری دارند.

سوال ۲:

برای یافتن اشکالات Multi-Schema و Multi-Instance نیاز به دو تا دیتاست داریم که چون نداریم پس مشکلی هم نمیتونیم پیدا کنیم.

Single-Instance:

در دیتاست وقتی validity رو چک میکنیم متوجه یک الگو بین patient id ها میشویم که شامل ۳ کاراکتر و ۳ حرف میباشند. وقتی با این الگو validity را چک میکنیم میبینیم یک سری از ایدی ها از این الگو پیروی نکردن و طبق اسلایدها مشکل instance دارن.

Single-Schema:

مشکل Income:

درآمد هایی که در دیتاست داریم با واحد مشخصی ذکر نشدن.(مثلا درآمد به دلار هست یا یورو)

و همچنین در کشورهای مختلف واحد درآمدها فرق میکند و ما نمیدانیم که شخص در کشور خودش به همان واحد درآمد دارد یا خیر.

مشکل blood pressure: داده ها باید عددی باشند و با این فرمت کاربر میتونه اشتباه هم وارد کنه.

در مورد violated attribute dependencies هم بین کشور و قاره داریم.

سوال۳:

بررسی و حذف داده های ناقص: ابتدا داد های دیتاست را بررسی کنیم و هر نوع داده ناقص را تشخیص داده و حذف کنیم. داده های ناقص ممکن است به دلیل عدم وجود مقدار یا خطا در سینتکس یا دلایل دیگر باشد.

پرکردن داده‌های خالی: در صورتی که داده‌های خالی یا ناقصی در دیتاست وجود داشته باشد، می‌توانیم از روش‌های پرکردن داده‌های خالی مانند میانگین‌گیری، حدس زدن مقدار، یا استفاده از روش‌های پیش‌بینی استفاده کنیم.

بررسی توزیع داده‌ها: بررسی توزیع داده‌ها و اطمینان حاصل کردن از اینکه داده‌ها به درستی توزیع شده‌اند. می‌توانیم از روش‌های تبدیل داده مانند استانداردسازی (standardization) یا نرمال‌سازی (normalization) استفاده کنیم

به‌روزرسانی داده‌ها: اگر دیتاست مبتنی بر اطلاعات گذشته است و از آخرین داده‌ها برخوردار نیست، می‌توانیم داده‌های جدید را به دیتاست اضافه کنیم تا نتایج پیش‌بینی بهتری ارائه شود.

میتوانیم یک دیتاست دیگه درست کنیم برای inconsistency بین کشور و قاره ها رو میشه از این طریق رفع کرد.