

«به نام خدا»

تکلیف ششم – سوال سوم – مرضیه علیدادی – 9631983
(کد های مربوط، در دو فرمت py و ipynb. ضمیمه شده اند.)

3.

(a)

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst radius	worst texture	worst perimeter	worst area	worst smoothness
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710	0.2419	0.07871	...	25.380	17.33	184.60	2019.0	0.16220
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	0.1812	0.05667	...	24.990	23.41	158.80	1956.0	0.12380
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790	0.2069	0.05999	...	23.570	25.53	152.50	1709.0	0.14440
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520	0.2597	0.09744	...	14.910	26.50	98.87	567.7	0.20980
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	0.1809	0.05883	...	22.540	16.67	152.20	1575.0	0.13740
...
564	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	0.1726	0.05623	...	25.450	26.40	166.10	2027.0	0.14100
565	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	0.1752	0.05533	...	23.690	38.25	155.00	1731.0	0.11660
566	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	0.1590	0.05648	...	18.980	34.12	126.70	1124.0	0.11390
567	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	0.2397	0.07016	...	25.740	39.42	184.60	1821.0	0.16500
568	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	0.1587	0.05884	...	9.456	30.37	59.16	268.6	0.08996

569 rows × 30 columns

(b)

]:

mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst texture	worst perimeter	worst area	worst smoothness	worst compactness	worst concavity	worst concave points	worst symmetry	worst fractal dimension	Cancer
0.27760	0.30010	0.14710	0.2419	0.07871	...	17.33	184.60	2019.0	0.16220	0.66560	0.7119	0.2654	0.4601	0.11890	0
0.07864	0.08690	0.07017	0.1812	0.05667	...	23.41	158.80	1956.0	0.12380	0.18660	0.2416	0.1860	0.2750	0.08902	0
0.15990	0.19740	0.12790	0.2069	0.05999	...	25.53	152.50	1709.0	0.14440	0.42450	0.4504	0.2430	0.3613	0.08758	0
0.28390	0.24140	0.10520	0.2597	0.09744	...	26.50	98.87	567.7	0.20980	0.86630	0.6869	0.2575	0.6638	0.17300	0
0.13280	0.19800	0.10430	0.1809	0.05883	...	16.67	152.20	1575.0	0.13740	0.20500	0.4000	0.1625	0.2364	0.07678	0
...
0.11590	0.24390	0.13890	0.1726	0.05623	...	26.40	166.10	2027.0	0.14100	0.21130	0.4107	0.2216	0.2060	0.07115	0
0.10340	0.14400	0.09791	0.1752	0.05533	...	38.25	155.00	1731.0	0.11660	0.19220	0.3215	0.1628	0.2572	0.06637	0
0.10230	0.09251	0.05302	0.1590	0.05648	...	34.12	126.70	1124.0	0.11390	0.30940	0.3403	0.1418	0.2218	0.07820	0
0.27700	0.35140	0.15200	0.2397	0.07016	...	39.42	184.60	1821.0	0.16500	0.86810	0.9387	0.2650	0.4087	0.12400	0
0.04362	0.00000	0.00000	0.1587	0.05884	...	30.37	59.16	268.6	0.08996	0.06444	0.0000	0.0000	0.2871	0.07039	1

(c) هیچ مقدار null ای در دیتاست وجود ندارد. به غیر از متغیر هدف، بقیه از نوع داده های عددی هستند. آن ها را نرمال سازی کردم:

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst texture	worst perimeter	worst area	w
0	0.007925	0.004573	0.054099	0.440986	0.000052	0.000122	0.000132	0.000065	0.000107	0.000035	...	0.007635	0.081325	0.889462	0.001
1	0.008666	0.007486	0.055988	0.558619	0.000036	0.000033	0.000037	0.000030	0.000076	0.000024	...	0.009862	0.066899	0.824026	0.001
2	0.009367	0.010109	0.061842	0.572276	0.000052	0.000076	0.000094	0.000061	0.000098	0.000029	...	0.012145	0.072545	0.812984	0.001
3	0.016325	0.029133	0.110899	0.551922	0.000204	0.000406	0.000345	0.000150	0.000371	0.000139	...	0.037881	0.141333	0.811515	0.001
4	0.009883	0.006985	0.065808	0.631774	0.000049	0.000065	0.000096	0.000051	0.000088	0.000029	...	0.008120	0.074137	0.767189	0.001
...
564	0.008541	0.008870	0.056256	0.585933	0.000044	0.000046	0.000097	0.000055	0.000068	0.000022	...	0.010459	0.065804	0.803033	0.001
565	0.009344	0.013114	0.060903	0.585355	0.000045	0.000048	0.000067	0.000045	0.000081	0.000026	...	0.017756	0.071951	0.803528	0.001
566	0.011644	0.019696	0.075966	0.601905	0.000059	0.000072	0.000065	0.000037	0.000112	0.000040	...	0.023933	0.088872	0.788417	0.001
567	0.009230	0.013142	0.062774	0.566806	0.000053	0.000124	0.000157	0.000068	0.000107	0.000031	...	0.017663	0.082713	0.815932	0.001
568	0.023110	0.073082	0.142710	0.539032	0.000157	0.000130	0.000000	0.000000	0.000473	0.000175	...	0.090444	0.176183	0.799912	0.001

569 rows × 31 columns

(f) دقت مدل برابر 94% است:

```
[[41  7]
 [ 0 66]]
```

	precision	recall	f1-score	support
0	1.00	0.85	0.92	48
1	0.90	1.00	0.95	66
accuracy			0.94	114
macro avg	0.95	0.93	0.94	114
weighted avg	0.94	0.94	0.94	114

داده هایی که در دسته ی 0 قرار می گیرند، 48 تا هستند. که 41 مورد از آن ها درست پیشبینی شده اند. داده هایی که در دسته ی 1 قرار می گیرند، 66 تا هستند. که همه ی آن ها درست پیشبینی شده اند.

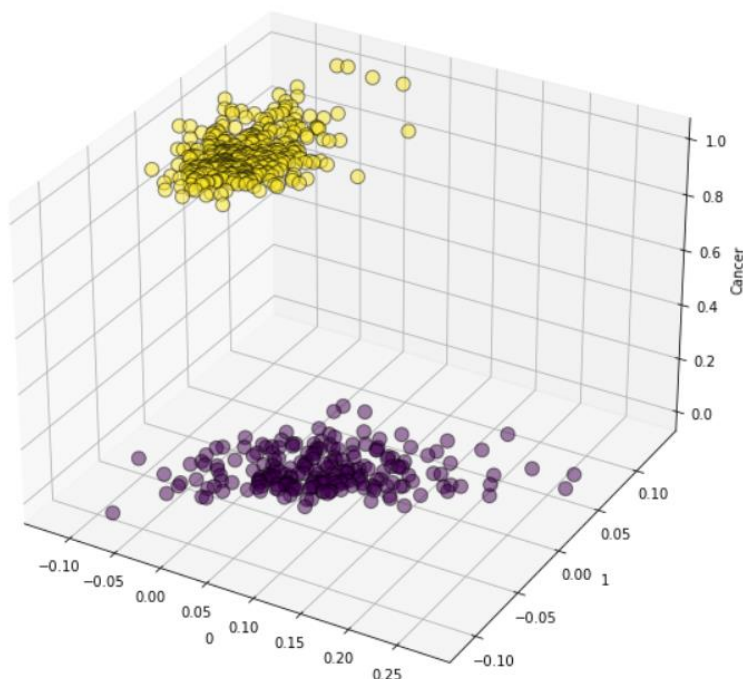
(g) پوشش متغیر ها بالاست:

```
array([0.76709045, 0.94500855])
```

و correlation آن ها با یکدیگر پایین است:

	0	1
0	1.000000e+00	2.962858e-17
1	2.962858e-17	1.000000e+00

نمودار:



(h) دقت مدل برابر 90% است:

```
[[39  9]
 [ 2 64]]
```

	precision	recall	f1-score	support
0	0.95	0.81	0.88	48
1	0.88	0.97	0.92	66
accuracy			0.90	114
macro avg	0.91	0.89	0.90	114
weighted avg	0.91	0.90	0.90	114

داده هایی که در دسته ی 0 قرار می گیرند، 48 تا هستند. که 39 مورد از آن ها درست پیشبینی شده اند.
داده هایی که در دسته ی 1 قرار می گیرند، 66 تا هستند. که 64 تا از آن ها درست پیشبینی شده اند.

(i) در حالت دوم که از PCA استفاده کردم، دقت مدل تقریباً 4% کاهش پیدا کرد.
احتمالاً دلیلش این است که، پوشش داده ها به 94% کاهش یافته است. و درحالت عادی، قبل از PCA، 100% بوده است.