«به نام خدا»

تکلیف دوم – سوال چهارم – مرضیه علیدادی – 9631983

(کد های مربوط، در دو فرمت py. و ipynb. ضمیمه شده اند. – دیتاست اصلاح شده Diabetes_cleared.csv نیز ضمیمه شده است.)

4.

a) نرمال سازی داده ها به معنی scale کردن آن ها در بازه های کوچک است. این کار برای الگوریتم های classification مفید است.

نرمال سازی زمانی مورد نیاز است که ما با متغیر هایی با رنج های متفاوت مواجه باشیم؛ چرا که در این صورت، الگوریتم، این گونه برداشت می کند که متغیری که رنج داده های وسیع تری دارد، اهمیت بیشتری نسبت به متغیر دیگر با رنج داده های کوچک تر دارد. اگرچه ممکن است این دو متغیر، از اهمیت یکسانی در تحلیل ما برخوردار باشند؛ یا حتی متغیری که مهم تر در نظر گرفته شده، کم تر اهمیت داشته باشد.

برای مثال در دیتاست مربوط به یک شرکت، دو متغیر "میزان تجربه" و "میزان حقوق" وجود دارد؛ که رنج اولی، از 0 تا 30 است و رنج دومی از 10000 تا 100000 است. می خواهیم بررسی کنیم که افراد در چه جایگاهی در آن شرکت قرار دارند. بنابراین، باید مدلی برای این کار با استفاده از الگوریتم ها تهیه کنیم. با توجه به این که بازه ی "میزان حقوق" گسترده تر است، اهمیت بیشتری برای آن در مدل در نظر گرفته می شود، نسبت به "میزان تجربه".

b) تحلیل ها در سوال 3 صورت گرفته است.

c)

- استاندارد سازی با MinMaxScaler :

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	0.352941	0.172589	0.456522	0.500000	0.158048	0.314928	0.234415	0.500000	1.0
1	0.058824	0.092640	0.391304	0.392857	0.158048	0.171779	0.116567	0.306122	0.0
2	0.470588	0.217005	0.369565	0.391152	0.158048	0.104294	0.253629	0.316327	1.0
3	0.058824	0.097716	0.391304	0.285714	0.084567	0.202454	0.038002	0.204082	0.0
4	0.000000	0.158629	0.108696	0.500000	0.162791	0.509202	0.943638	0.326531	1.0
5	0.294118	0.131980	0.478261	0.391152	0.158048	0.151329	0.052519	0.295918	0.0
6	0.176471	0.083756	0.217391	0.446429	0.078224	0.261759	0.072588	0.255102	1.0
7	0.588235	0.130711	0.459399	0.391152	0.158048	0.349693	0.023911	0.285714	0.0
8	0.117647	0.234772	0.434783	0.678571	0.559197	0.251534	0.034159	0.530612	1.0
9	0.470588	0.143401	0.717391	0.391152	0.158048	0.292587	0.065756	0.540816	1.0
10	0.235294	0.124365	0.673913	0.391152	0.158048	0.396728	0.048249	0.295918	0.0
11	0.588235	0.197970	0.478261	0.391152	0.158048	0.404908	0.195986	0.336735	1.0
12	0.588235	0.161168	0.543478	0.391152	0.158048	0.182004	0.581981	0.571429	0.0
13	0.058824	0.224619	0.326087	0.285714	0.879493	0.243354	0.136635	0.591837	1.0
14	0.294118	0.195431	0.456522	0.214286	0.170190	0.155419	0.217336	0.510204	1.0

- استاندارد سازی با StandardScaler :

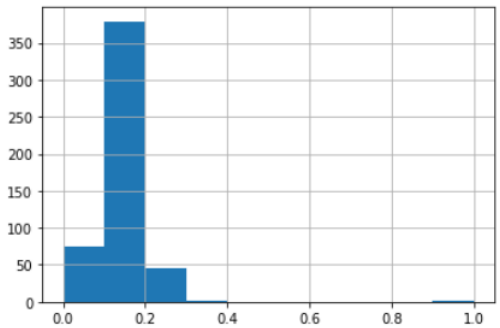
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	0.658827	0.541810	-2.248873e-02	0.710169	0.000000	0.153840	0.410565	1.402872	1.345423
1	-0.840407	-0.886988	-5.322332e-01	0.011127	0.000000	-0.831880	-0.390056	-0.164487	-0.750738
2	1.258521	1.335587	-7.021480e-01	0.000000	0.000000	-1.296577	0.541101	-0.081995	1.345423
3	-0.840407	-0.796270	-5.322332e-01	-0.687915	-0.673741	-0.620655	-0.923804	-0.989413	-0.750738
4	-1.140254	0.292338	-2.741126e+00	0.710169	0.043484	1.491603	5.228794	0.000498	1.345423
5	0.358980	-0.183928	1.474261e-01	0.000000	0.000000	-0.972698	-0.825176	-0.246980	-0.750738
6	-0.240714	-1.045743	-1.891552e+00	0.360648	-0.731894	-0.212285	-0.688839	-0.576950	1.345423
7	1.858214	-0.206608	-1.207317e-15	0.000000	0.000000	0.393229	-1.019530	-0.329472	-0.750738
8	-0.540561	1.653097	-1.924035e-01	1.875240	3.678068	-0.282693	-0.949911	1.650350	1.345423
9	1.258521	0.020186	2.016489e+00	0.000000	0.000000	0.000000	-0.735252	1.732842	1.345423
10	0.059133	-0.320004	1.676660e+00	0.000000	0.000000	0.717109	-0.854184	-0.246980	-0.750738
11	1.858214	0.995397	1.474261e-01	0.000000	0.000000	0.773436	0.149493	0.082991	1.345423
12	1.858214	0.337696	6.571706e-01	0.000000	0.000000	-0.761472	2.771816	1.980320	-0.750738
13	-0.840407	1.471663	-1.041978e+00	-0.687915	6.614812	-0.339020	-0.253719	2.145305	1.345423
14	0.358980	0.950038	-2.248873e-02	-1.153944	0.111330	-0.944534	0.294533	1.485365	1.345423

- استاندارد سازی با Normalize :

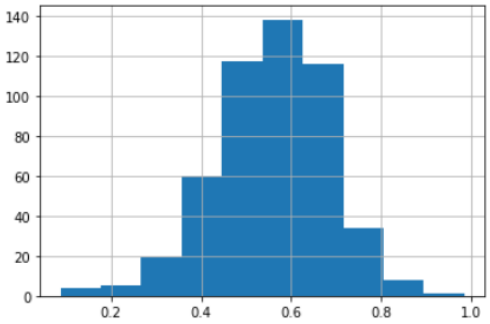
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	0.024761	0.610778	0.297135	0.144441	0.674800	0.138663	0.002588	0.206344	0.004127
1	0.004949	0.420667	0.326635	0.143522	0.809231	0.131644	0.001737	0.153420	0.000000
2	0.030955	0.708104	0.247643	0.111844	0.632703	0.090158	0.002600	0.123821	0.003869
3	0.006612	0.588467	0.436392	0.152076	0.621527	0.185797	0.001104	0.138852	0.000000
4	0.000000	0.596381	0.174126	0.152360	0.731328	0.187620	0.009960	0.143654	0.004353
5	0.022802	0.528999	0.337465	0.131814	0.745677	0.116745	0.000917	0.136810	0.000000
6	0.021765	0.565885	0.362747	0.232158	0.638435	0.224903	0.001799	0.188628	0.007255
7	0.045546	0.523776	0.329135	0.131648	0.744734	0.160776	0.000610	0.132083	0.000000
8	0.003408	0.335723	0.119292	0.076688	0.925367	0.051977	0.000269	0.090321	0.001704
9	0.033671	0.526107	0.404051	0.121655	0.688205	0.136820	0.000976	0.227278	0.004209
10	0.017806	0.489659	0.409533	0.128667	0.727871	0.167374	0.000850	0.133543	0.000000
11	0.039537	0.664216	0.292571	0.114279	0.646478	0.150239	0.002123	0.134425	0.003954
12	0.041747	0.580283	0.333976	0.120667	0.682619	0.113134	0.006016	0.237958	0.000000
13	0.001147	0.216804	0.068827	0.026384	0.970457	0.034528	0.000457	0.067680	0.001147
14	0.019314	0.641223	0.278121	0.073393	0.675988	0.099660	0.002267	0.197002	0.003863

d)

- نمودار حاصل از استاندارد سازی با MinMaxScaler :



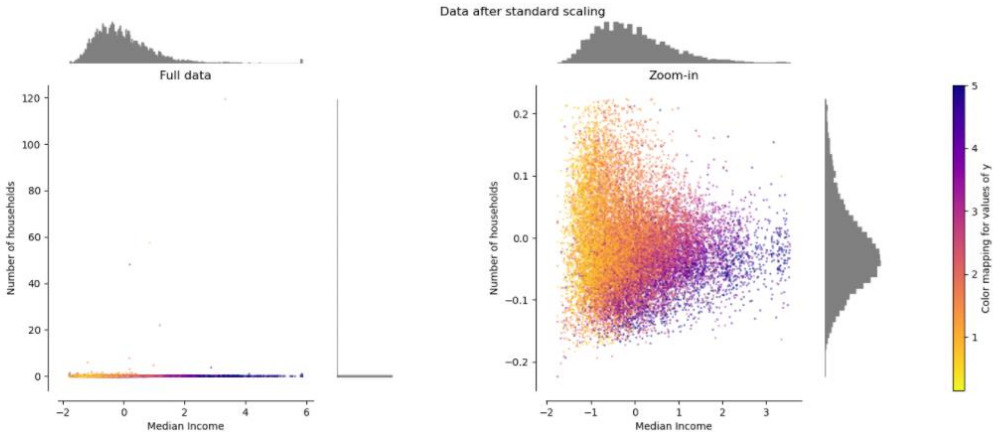
- نمودار حاصل از استاندارد سازی با Normalize:



در نمودار مربوط به دیتای نرمال شـده با روش MinMaxScaler، کجی راسـت مشـاهده می شـود. ولی نمودار مربوط به دیتای نرمال شده با روش Normalize، نسبتاً به توزیع نرمال نزدیک تر است.

e)

- StandardScaler میانگین را حذف می کند و داده ها را به واحد واریانس مقیاس بندی می کند. این مقیاس بندی، دامنه ی مقادیر متغیر ها را کاهش می دهد. اگرچه، داده های پرت، روی محاسبه ی میانگین و انحراف معیار تأثیر دارد. با توجه به اینکه داده های پرت برای هر کدام از متغیر ها دارای اندازه ی متفاوتی هستند، گسترش داده های نرمال شـــده با این روش، برای هر کدام از متغیر ها متفاوت اســـت. بنابراین، این روش در حضـــور داده های پرت، نمی تواند متعادل بودن مقیاس متغیر ها را تضمین کند. این روش روی ستون اجرا می شود.



- Normalize مستقل از توزیع نمونه، مقادیر هر سمپل را به فرم یکسان می برد. این روش روی سطر اجرا می شود.