به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس سیستم‌های هوشمند

تمرین شماره ۴

نام و نام خانوادگی : سیاوش شمس

شماره دانشجویی : ۸۱۰۱۹۷۶۴۴

دی ۱۴۰۰

فهرست سوالات

سوال ۱ ..... ۳

الف: ۳

ب: ۵

سوال ۲ ..... ۸

الف: ۸

ب: ۱۱

سوال ۳ ..... ۱۴

الف: ۱۴

ب: ۱۶

ج: ۱۶

د: ۱۸

سوال ۴ ..... ۲۴

الف-۱: ۲۴

الف-۲: ۲۵

الف۲-۱: ۲۵

الف۲-۲: ۲۶

الف-۳: ۲۶

ب-۱: ۲۷

ب-۲: ۲۸

پیوست ۳۱

سوال ۱

در قسمت اول این سوال داده های داده شده را به کمک الگوریتم کا-میانگین [1]دسته بندی می کنیم، در قسمت دوم این سوال داده های داده شده را به کمک الگوریتم خوشه بندی سلسله مراتبی[2] و معیار پیوند واحد [3]دسته بندی می کنیم.

الف:

به طور تصادفی دو نقطه B و C را به عنوان نقطه مرکزی ابتدایی در نظر می گیریم. فاصله ها را حساب می کنیم و آن ها را در جدول می نویسیم. اگر فاصله نقطه ای از B کمتر بود به آن برچسب ۱ را تحت ستون Cluster اختصاص می دهیم و اگر فاصله نقطه از C کمتر بود به آن برچسب ۲ را اختصاص می دهیم.

جدول ۱-۱- فاصله هر نقطه از مرکز خوشه و برچسب آن بر اساس فاصله تکرار ۱

distance	B (1)	C (2)	Cluster
A	1	2	1
B	0	5	1
C	5	0	2
D	17	8	2
E	29	18	2

حال مرکز هر دسته را با توجه به داده های برچسب خورده حساب می کنیم.

$$(1): mean(A, B) = (1, 0.5)$$

$$(2): mean(C, D, E) = (\frac{5}{3}, \frac{11}{3})$$

[1] K-means
[2] Hierarchical Clustering
[3] Single link

دوباره جدول فاصله ها را تشکیل می دهیم:

جدول ۱-۲- فاصله هر نقطه از مرکز خوشه و برچسب آن بر اساس فاصله تکرار ۲

distance	(1)	(2)	Cluster
A	0.25	7.56	1
B	0.25	13.89	1
C	3.25	5.56	1
D	13.25	0.22	2
E	24.25	3.56	2

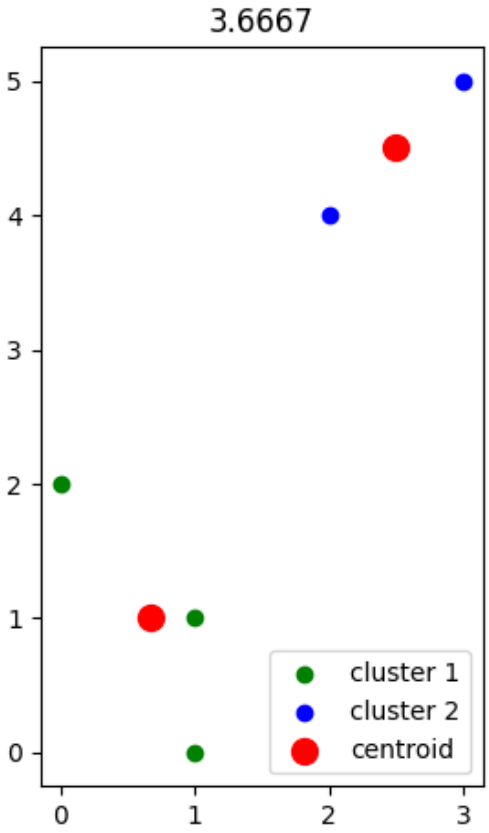
$$(1): mean(A, B) = \left(\frac{2}{3}, 1\right)$$

$$(2): mean(C, D, E) = (\frac{5}{2}, \frac{9}{2})$$

جدول ۱-۳- فاصله هر نقطه از مرکز خوشه و برچسب آن بر اساس فاصله تکرار ۳

distance	(1)	(2)	Cluster
A	0.11	14.5	1
B	1.11	22.5	1
C	1.44	12.5	1
D	10.78	0.5	2
E	21.44	0.5	2

با توجه به جدول ۱-۳ می بینیم که طبقه بندی عوض نشد پس به نقطه بهینه همگرا شدیم و الگوریتم به پایان می رسد.



شکل ۱-۱- داده ها و مرکز خوشه ها رسم شده به کمک پایتون

ب:

جدول فاصله هر نقطه از سایر نقاط را تشکیل می دهیم.

$$d(P1,P2) = \sqrt{(P1-P2)^2}$$

جدول ۱-۴- فاصله هر نقطه نقاط دیگر تکرار اول

	P1	P2	P3	P4	P5
P1	0	0.14	0.19	0.14	0.24
P2	0.14	0	0.16	0.28	0.1
P3	0.19	0.16	0	0.28	0.22
P4	0.14	0.28	0.28	0	0.39
P5	0.24	0.1	0.22	0.39	0

بر اساس جدول خوشه اول را متشکل از (P5,P2) انتخاب می کنیم. حال بر اساس روش پیوند واحد[1] فاصله های جدید نقاط را به دست می آوریم.

جدول ۱-۵- فاصله هر نقطه نقاط دیگر تکرار اول دوم

	P1	P2, P5	P3	P4
P1	0	0.14	0.19	0.14
P2, P5	0.14	0	0.16	0.28
P3	0.19	0.16	0	0.28
P4	0.14	0.28	0.28	0

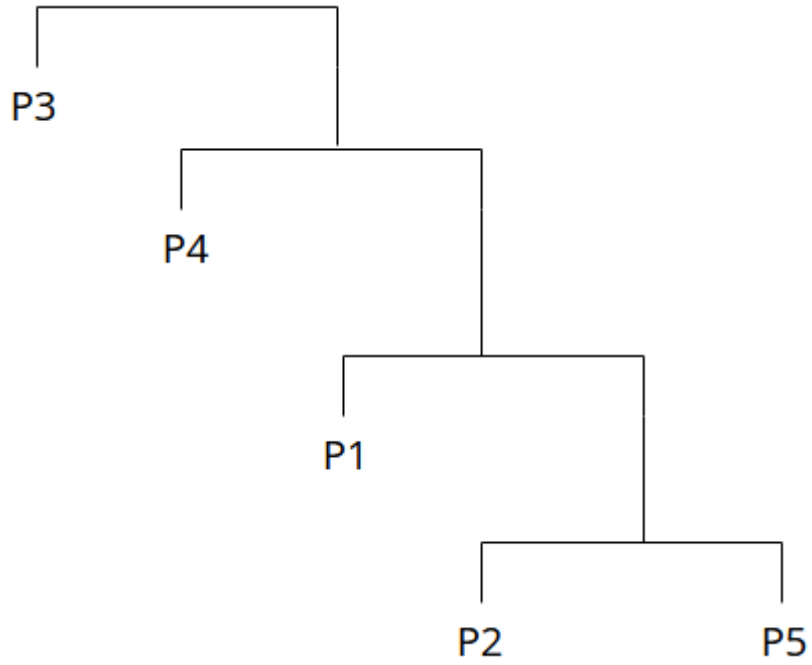
بر اساس جدول جدید، خوشه دوم را متشکل از ((P1,(P2,P5)) در نظر می گیریم.

جدول ۱-۶- فاصله هر نقطه نقاط دیگر تکرار سوم

	P1, P2, P5	P3	P4
P1, P2, P5	0	0.14	0.28
P3	0.14	0	0.28
P4	0.28	0.28	0

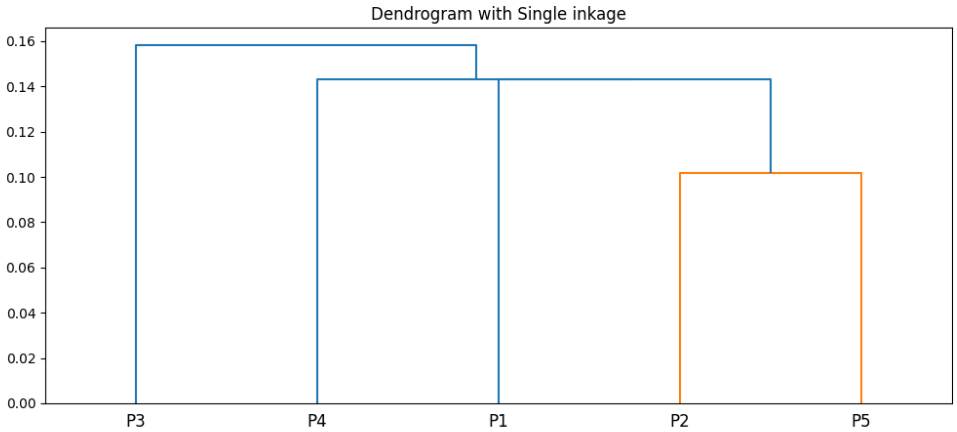
بر اساس جدول جدید، خوشه دوم را متشکل از ((P4,(P1,P2,P5)) در نظر می گیریم.

[1] Single link



شکل ۱-۲- نمودار درختی خوشه بندی رسم شده

همچنین این قسمت را برای اطمینان بیشتر به کمک پایتون پیاده سازی کردیم و نمودار درختی آن به صورت زیر در آمد که تفاوت آن در این است که به دلیل برابر بودن فاصله d(P4, P1) و d(P1,(P2,P5)) در جدول ۱-۵، دو نقطه P1 و P4 در یک مرحله خوشه بندی شده اند. کد این بخش در پیوست آمده است.



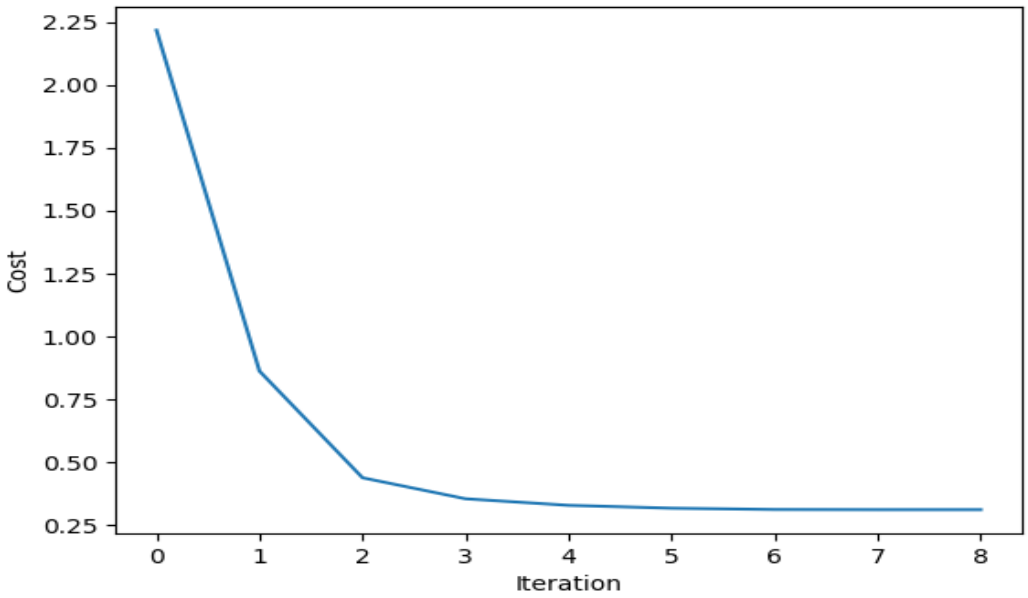
شکل ۱-۳- نمودار درختی خوشه بندی رسم شده به کمک پایتون

سوال ۲

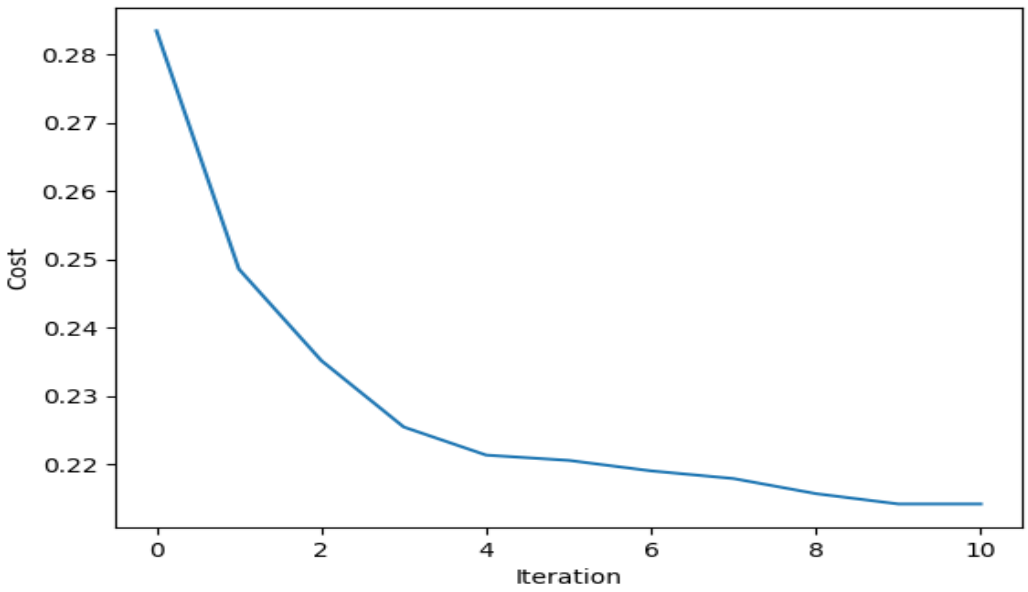
در قسمت اول این سوال الگوریتم کا-میانگین را به ازای K=5,10,20 پیاده سازی می کنیم، و تابع هزینه در هر مرحله تکرار درونی الگوریتم(تکرار ها قبل از رسیدن به نقطه بهینه) را رسم می کنیم، سپس نمودار تابع هزینه و معیار شباهت داخلی نسبت به شباهت بیرونی به ازای K بین ۲ تا ۲۰ را رسم می کنیم و سعی می کنیم نقطه زانویی نمودار را پیدا کنیم که به ما مقدار مناسب K را نشان می دهد. در قسمت بعد الگوریتم خوشه بندی خود را ۱۵ بار با مرکز های اولیه جدید آغاز کرده و میانگین و واریانس مربوط به هر خوشه را در هر تکرار نمایش می دهیم، در نهایت میانگین و واریانس برای معیار شباهت درونی و بیرونی که از تکرار های مختلف به دست می آیند را حساب می کنیم و با هم مقایسه می کنیم.

الف:

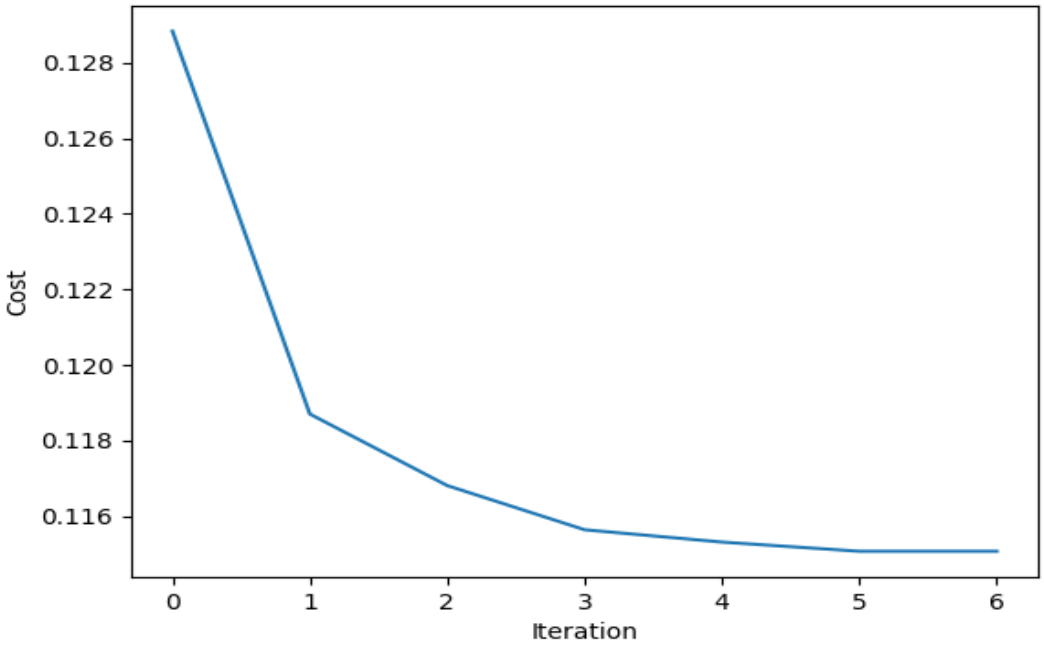
در نمودار های زیر مقدار تابع هزینه تا وقتی که الگوریتم به نقطه بهینه رسیده است رسم شده است.



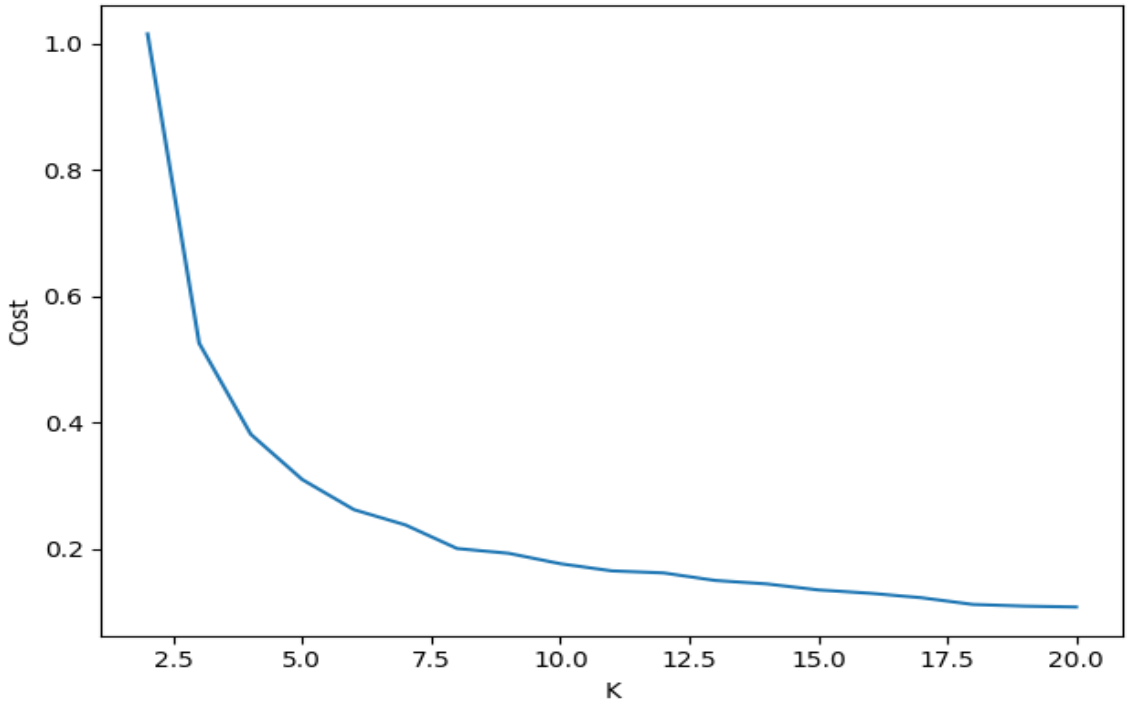
شکل ۲-۱- نمودار مقدار تابع هزینه به ازای هر تکرار درونی الگوریتم در K=5



شکل ۲-۲ نمودار مقدار تابع هزینه به ازای هر تکرار درونی الگوریتم در K=10



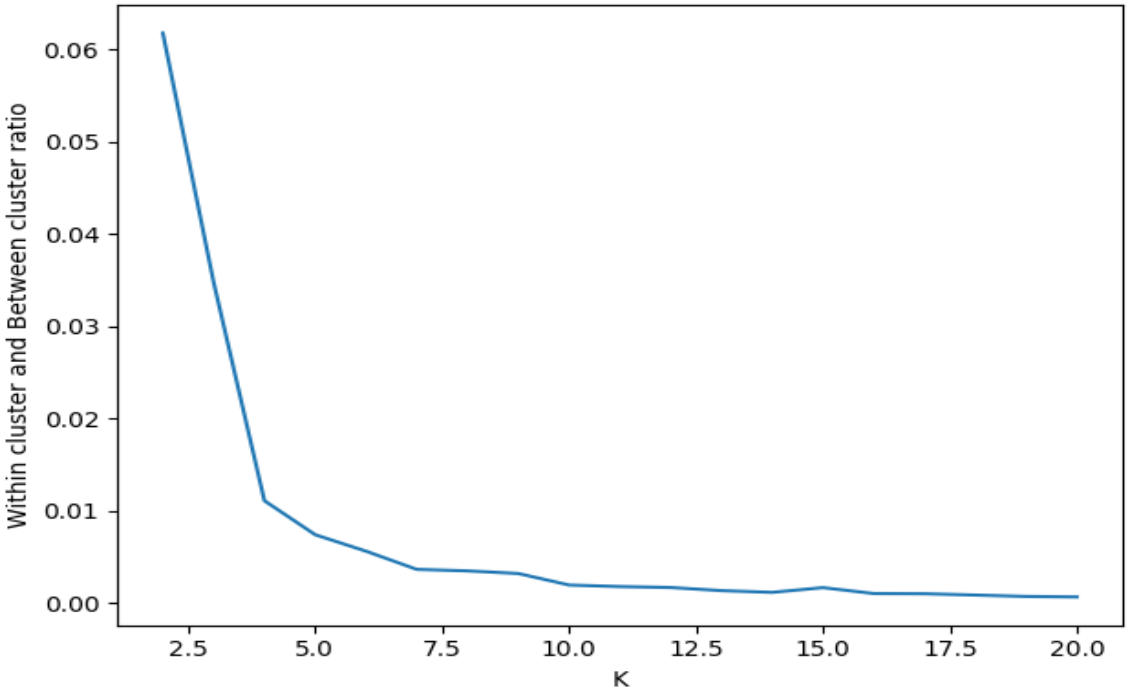
شکل ۲-۳- نمودار مقدار تابع هزینه به ازای هر تکرار درونی الگوریتم در K=20



شکل ۲-۴- نمودار مقدار تابع هزینه به ازای K های مختلف

با توجه به نمودار ۲-۴ که در آن ها هزینه به ازای K های مختلف رسم شده است، می بینیم K=3 تا K=8 تقریبا نقطه زانویی نمودار هستند و می توانند به عنوان K بهینه انتخاب شوند.

همچنین نمودار معیار پیشنهادی در صورت سوال به ازای K های مختلف را می کشیم.



شکل ۲-۵- نمودار مقدار نسبت شباهت درونی به بیرونی به ازای K های مختلف

با توجه به نمودار ۲-۵ نیز می توان نتیجه گرفت K=3 یا K=5 می تواند مقدار خوبی برای K باشد.

ب:

جدول ۲-۱- واریانس هر خوشه در ۱۰ تکرار مختلف به ازای K=5

	0	1	2	3	4
0	2.91466	4.21694	2.97285	3.91393	3.37917
1	3.25295	3.39929	2.53903	3.94742	2.69217
2	3.90265	2.91466	4.21694	3.37917	2.99632
3	3.22981	2.94535	4.23309	2.68626	3.71285
4	3.71285	3.94742	2.62328	2.94535	3.24477
5	3.22531	2.62812	3.94036	3.94742	3.08754
6	3.00750	3.39929	2.62328	4.31000	3.52487
7	4.09082	3.08754	2.62328	3.23968	3.94036
8	2.62496	5.03651	3.23658	3.39929	3.19166
9	3.19166	3.27529	3.39929	2.62328	5.03651

جدول ۲-۲- واریانس هر خوشه در ۱۰ تکرار مختلف به ازای K=10

0	5.03651	2.89943	2.61567	3.60069	3.04109	4.29790	3.32000	3.03056	3.22694	3.19166
1	2.87562	3.70822	3.71285	5.03651	3.61243	2.45114	3.91687	2.75331	2.94535	3.23481
2	2.38910	2.12500	3.09266	2.71873	2.85674	2.73585	3.52050	3.52636	5.03651	3.39929
3	2.65569	2.94679	4.04259	3.22510	3.55484	5.03651	2.79335	2.53903	3.00035	3.36332
4	3.00027	5.03651	3.64080	3.39929	2.63210	2.44403	2.90590	3.23698	3.63127	3.61243
5	2.56674	2.53903	3.72151	3.71285	2.64375	3.20406	3.32193	3.46582	5.26549	2.94535
6	3.66910	3.03056	3.16160	5.11669	2.60018	4.29790	2.66497	2.91466	3.54124	3.60069
7	2.67114	2.12500	4.31000	3.37917	4.21694	2.91466	3.00750	2.38910	2.82320	3.52487
8	4.04259	2.59362	2.42000	3.40654	2.94679	2.59859	3.39493	5.11669	3.33625	2.98167
9	3.81531	2.12500	4.29548	4.29937	2.66880	3.19166	3.31068	3.27529	5.03651	2.91466

جدول ۲-۳ واریانس هر خوشه در ۱۰ تکرار مختلف به ازای K=20

	0	1	2	3	4	5	6	7	8	9	10	1
0	5.39667	2.64610	3.60682	3.04316	3.98818	2.94833	3.43684	4.62354	3.35632	3.57437	2.56490	2.067
1	3.44309	2.53965	3.43187	2.83194	3.09718	2.06743	2.29688	3.06716	2.46934	2.44188	3.38726	5.036
2	2.24688	3.70822	3.37824	3.21500	3.04316	2.61490	4.29790	2.77856	3.39924	2.81187	2.54832	2.915
3	4.82160	2.88810	3.69866	2.42000	3.57437	3.85076	2.74109	2.82803	2.76323	2.12500	4.42760	3.429
4	5.49287	3.34238	4.04259	4.84318	2.06187	2.58434	2.14576	3.36790	3.59582	3.37417	3.30052	4.167
5	4.75972	3.16986	3.06672	2.77559	2.90490	2.60937	5.13209	2.94679	3.26750	3.32707	3.72151	2.214
6	2.53965	4.67354	3.59562	3.70822	3.95871	3.57437	5.11669	3.13165	3.26609	2.48961	3.04109	2.866
7	2.98647	2.31748	2.76900	3.32095	2.53965	4.60383	2.42168	3.90715	2.45326	4.08647	3.61243	3.916
8	3.04250	3.48963	3.76887	3.02240	3.02953	2.84986	4.55812	2.45609	3.10722	2.12500	3.72151	2.788
9	2.16277	3.24168	2.70908	3.09266	2.90590	5.13209	3.28040	2.71276	3.94290	3.79294	2.96105	3.069

جدول۲-۴- میانگین هر خوشه در ۱۰ تکرار مختلف به ازای K=5

	0	1	2	3	4
0	2.33553	2.78250	3.61189	4.42500	2.59881
1	3.89671	2.53550	3.15192	4.49766	3.52500
2	4.43421	2.33553	2.78250	2.59881	3.61935
3	3.96862	2.36023	4.59239	3.35833	2.67321
4	2.67321	4.49766	3.33929	2.36023	3.88687
5	3.88110	3.32778	2.74118	4.49766	2.42955
6	4.04914	2.53550	3.33929	4.60000	3.84643
7	4.54630	2.42955	3.33929	3.92556	2.74118
8	3.31538	4.73750	3.85000	2.53550	4.31458
9	4.31458	3.86111	2.53550	3.33929	4.73750

جدول۲-۵- میانگین هر خوشه در ۱۰ تکرار مختلف به ازای K=10

	0	1	2	3	4	5	6	7	8	9
0	4.73750	2.32941	3.30500	2.59861	2.38750	2.83214	3.10000	2.63333	3.84295	4.31458
1	3.87647	3.93958	2.67321	4.73750	3.49167	3.13958	3.92500	3.56375	2.36023	4.32717
2	3.53000	2.90000	4.40333	3.28864	3.88281	3.50937	3.84875	4.15556	4.73750	2.53550
3	3.51562	3.89583	2.76731	4.33864	3.85526	4.73750	2.29062	3.15192	2.36818	2.57917
4	3.89375	4.73750	3.85556	2.53550	3.49063	3.12273	3.83125	4.34500	4.16500	3.49167
5	3.89750	3.15192	3.95909	2.67321	3.52500	4.45156	3.73750	4.12083	4.74750	2.36023
6	4.18000	2.63333	4.41094	4.79250	3.78472	2.83214	3.28804	2.33553	3.86304	2.59861
7	3.46042	2.90000	4.60000	2.59881	2.78250	2.33553	4.04914	3.53000	3.24643	3.84643
8	2.76731	3.29479	2.10000	2.56111	3.89583	2.21250	3.80833	4.79250	4.34896	2.41071
9	2.64286	2.90000	2.85500	2.77500	3.41250	4.31458	2.59265	3.86111	4.73750	2.33553

جدول۲-۶- میانگین هر خوشه در ۱۰ تکرار مختلف به ازای K=20

		9	10	11	12	13	14	15	16	17	18	19
0	75	4.15313	2.19000	2.90833	2.42500	2.87500	3.29792	2.87500	4.41786	2.65833	3.56607	3.99773
1	75	3.82500	3.86563	4.73750	2.35000	4.18611	2.81563	2.65625	3.86429	4.04500	3.65769	3.24583
2	64	3.37500	3.25417	2.48125	4.73750	2.65357	2.90000	4.23462	3.36667	4.44750	2.19000	2.42500
3	44	2.90000	2.82000	2.51154	4.26875	3.27500	4.56250	3.86000	2.62500	3.87885	3.27500	4.90000
4	33	3.85000	4.20556	3.95000	2.33194	3.70000	3.50625	4.06875	4.06250	3.93333	3.36528	3.88000
5	00	3.47083	3.95909	3.32500	3.52500	2.59375	2.92500	2.89167	2.19000	2.77917	2.42083	2.50000
6	60	3.16964	2.38750	2.35000	2.80000	4.46136	3.88281	2.64167	2.55500	2.87500	3.58295	4.29643
7	67	2.71786	3.49167	2.78750	4.27917	3.83750	3.88333	4.90000	3.45833	3.89500	2.35000	4.43333
8	33	2.90000	3.95909	2.29722	3.90500	2.49167	4.25750	4.79250	3.66250	2.74000	2.10000	3.73750
9	7	3.95750	3.88269	2.39643	2.99167	2.82500	4.52000	2.10000	4.15556	2.51250	2.45000	2.57500

جدول۲-۷- میانگین و واریانس معیار شباهت درونی و بیرونی به ازای K های مختلف در ۱۵ تکرار

	5	10	20
mean	0.0076184440655530355	0.0021672622609395023	0.000669459916363342
var	6.82588759908168e-07	6.302604644957808e-08	2.5871529611864583e-09

با توجه به جداول بالا می بینیم که با زیاد شدن تعداد K به طور کلی واریانس کاهش می یابد. دلیل آن نزدیک شدن داده های داخل هر خوشه به هم است، همچنین با افزایش K به طور کلی تفاوت بین میانگین های خوشه های مختلف کمتر می شود که با شهودی که از این روش دسته بندی داریم مطابقت دارد. نکته دیگر این است که میانگین و واریناس هر خوشه در هر تکرار متفاوت است که به دلیل تصادفی انتخاب شدن مرکز خوشه های اولیه می باشد.

سوال ۳

در قسمت اول این سوال ابتدا رگرسیون لجستیک [1]را مختصرا توضیح می دهیم سپس نمودار هیستوگرام داده های آموزش (۷۵٪ داده ها) را بر اساس برچسب آن ها رسم می کنیم، در قسمت ب یک طبقه بند رگرسیون لجستیک را با داده های آموزش برچسب دار آموزش می دهیم و داده های تست را با توجه به آن پیش بینی می کنیم و ماتریس آشفتگی و دقت را گزارش می کنیم. در قسمت ج الگوریتم خود تعلیم [2]را پیاده سازی کرده و ماتریس آشفتگی و دقت و معیار F1 را با قسمت قبل مقایسه می کنیم. در قسمت د نمودار معیار F1 و تعداد داده برچسب زده شده را به ازای آستانه های مختلف می کشیم و نتیجه می گیریم کدام آستانه بیشترین بهبود را در مدل ما حاصل می کند.

الف:

رگرسیون لجستیک یک مدل آماری رگرسیون برای متغیرهای وابسته که در آن رخ داد یک واقعه تصادفی با دو حالت ممکن است می باشد. در این متغیر ها مجموع احتمال هر حالت ممکن در نهایت یک خواهد شد. برای مثال: بیماری یا سلامت، مرگ یا زندگی، از این نوع متغیر ها هستند.

تفاوت مهم رگرسیون لجستیک با رگرسیون خطی در دو ویژگی رگرسیون لجستیک می‌تواند دیده شود. اول توزیع شرطی $y|\vec{x}$ یک توزیع برنولی به جای یک توزیع گوسی است چونکه متغیر وابسته دودویی[3] است. دوم اینکه مقادیر پیش‌بینی احتمالاتی که بین بازه صفر و یک قرار دارند به کمک تابع توزیع لجستیک به دست می‌آید، به عبارت دیگر رگرسیون لجستیک احتمال خروجی پیش‌بینی می‌کند.

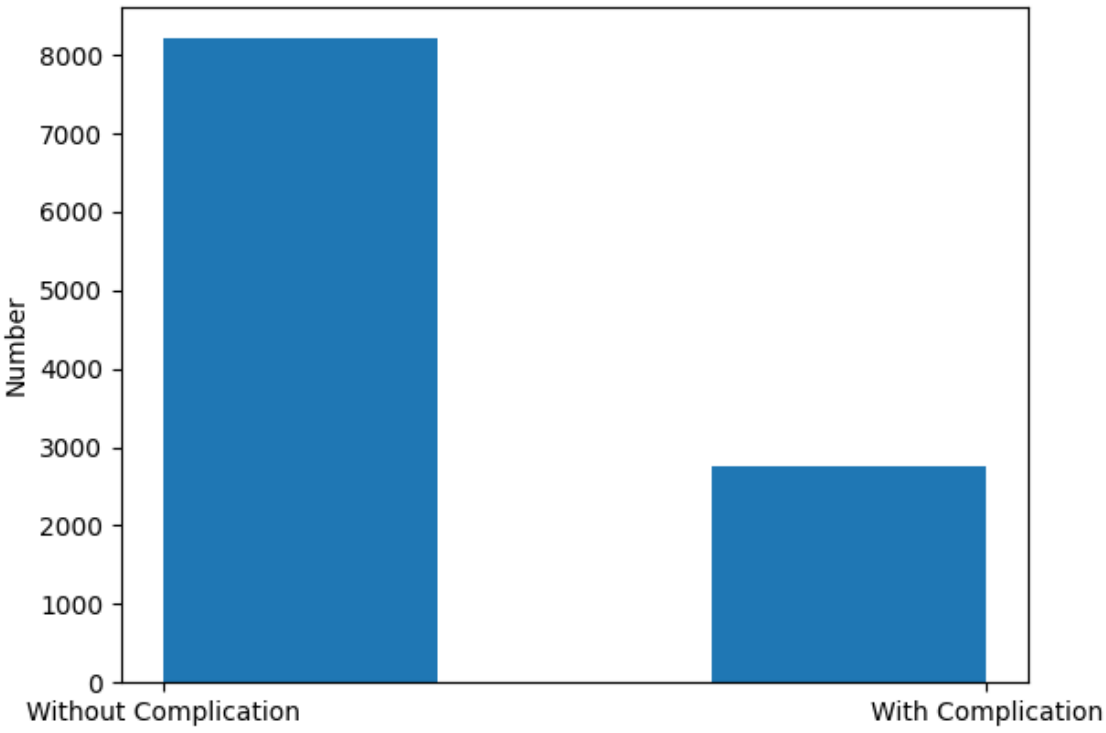
رگرسیون لجستیک را می‌توان به کمک تابع لجستیک تعریف کرد. دامنه این تابع اعداد حقیقی هستند و برد این تابع بین صفر و یک می‌باشد

$$\sigma = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}, \quad \sigma: \mathbb{R} \to [0,1]$$

با احتساب تابع لجستیک، رگرسیون لجستیک را می‌توان به شکل پایین نوشت:

$$\Pr(y_i = 1|x_i; \vec{\beta}) = \sigma(\vec{x}.\vec{\beta}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}}, \quad \sigma: \mathbb{R} \to [0,1]$$

[1] Logistic regression
[2] Self training
[3] Binary

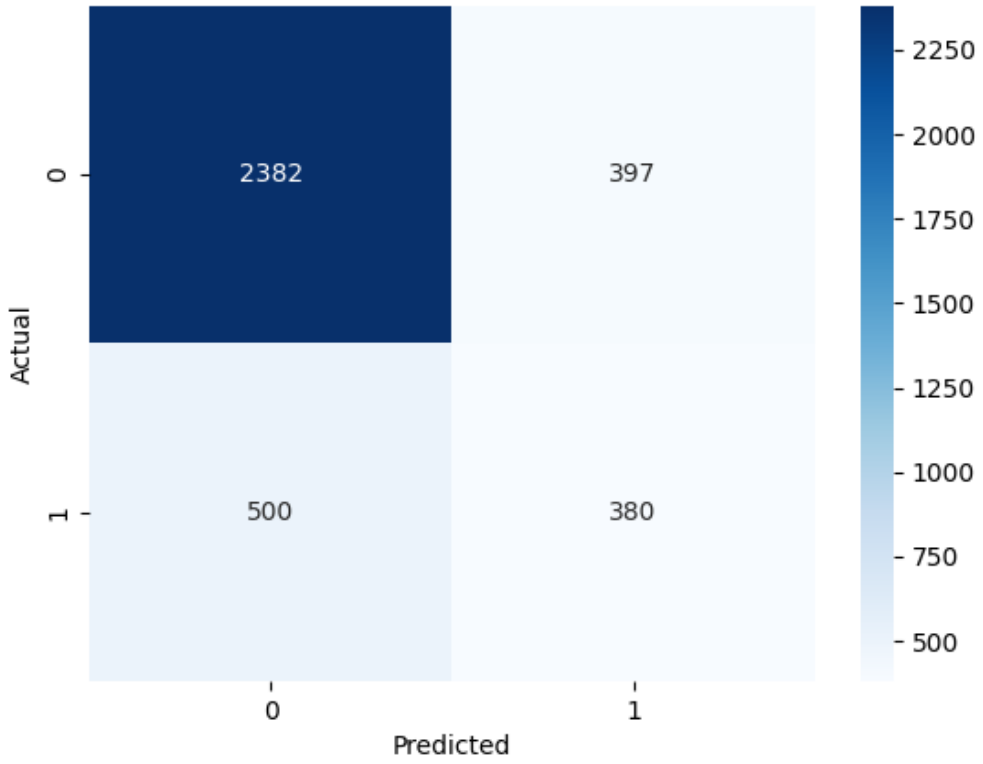


شکل ۳-۱- نمودار هیستوگرام توزیع داده های آموزش بر اساس برچسب آن ها

با توجه به نمودار شکل۳-۱ می بینیم که توزیع داده ها نا متعادل است و تعداد برچسب بدون عوارض[1] بیشتر از تعداد برچسب دارای عوارض می باشد.

[1] Complication

ب:



شکل ۳-۲- ماتریس آشفتگی عملکرد مدل لجستیک رگرسیون

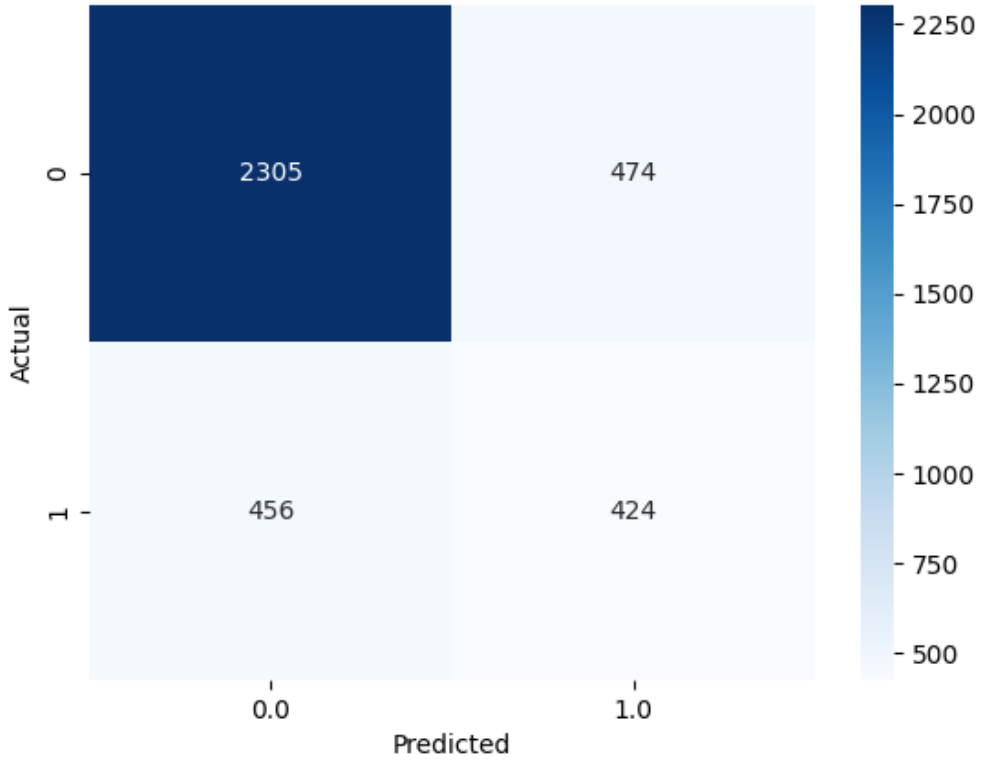
```
Accuracy of the model is:  0.7548510522000547
F1 score of the model is:  [0.84154743 0.45866023]
```

شکل ۳-۳- دقت و معیار F1 برای مدل

ج:

الگوریتم خود تعلیم در ۴ مرحله خلاصه می شود:

۱. ابتدا داده های دارای برچسب را به دو بخش آموزش و ارزیابی جدا می کنیم
۲. سپس مدل خود را با توجه به داده های برچسب دار آموزش می دهیم
۳. به کمک مدل آموزش داده شده برچسب احتمالی داده های بدون برچسب را پیش بینی می کنیم و داده هایی که با احتمال بیشتر از حدی (مثلا ۹۹٪) پیش بینی شده اند را جزو داده های آموزش می کنیم
۴. مراحل بالا را آن قدر انجام می دهیم تا دیگر داده بدون برچسبی قابل اضافه کردن نباشد.



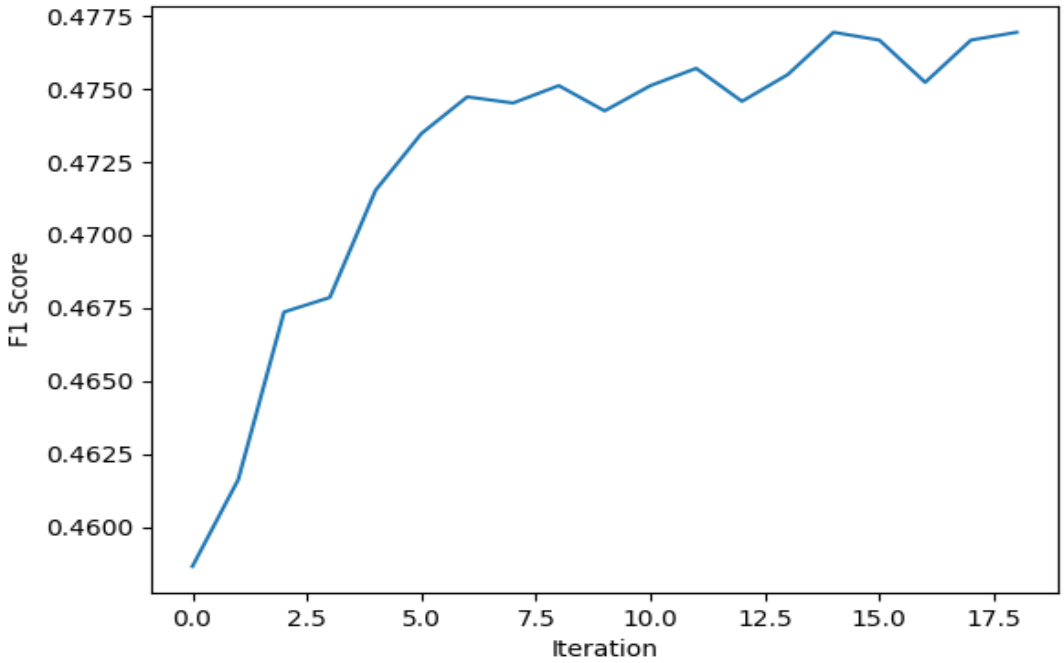
شکل ۳-۴- ماتریس آشفتگی عملکرد مدل پس از پیاده سازی الگوریتم خود تعلیم

```
Accuracy of the model is:  0.7458321945886854
F1 score of the model is:  [0.83212996 0.47694038]
```

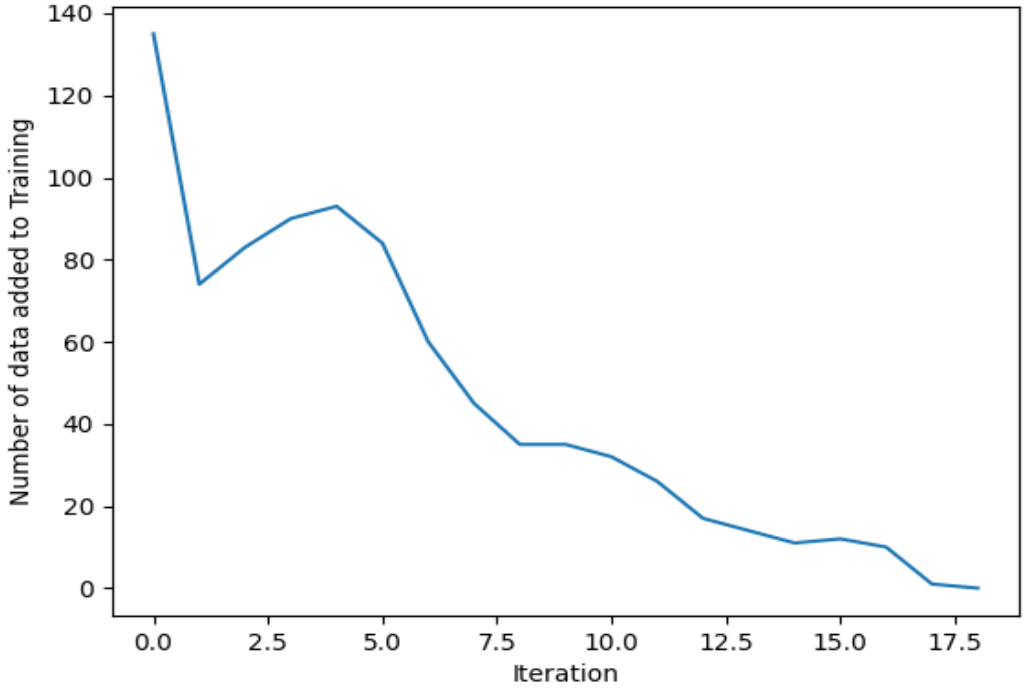
شکل ۳-۵- دقت و معیار F1 برای مدل

با توجه به شکل ۳-۵ می بینیم روش خود تعلیم دقت مدل ما را کمی کاهش داده اما معیار F1 برای داده های TP بیشتر شده است که در کاربرد های پزشکی این مقدار هم ارزشمند است. زیرا تشخیص درست بیمار بودن مهم تر از تشخیص درست بیمار نبودن است.

د:



شکل ۳-۶- معیار F1 به ازای هر تکرار



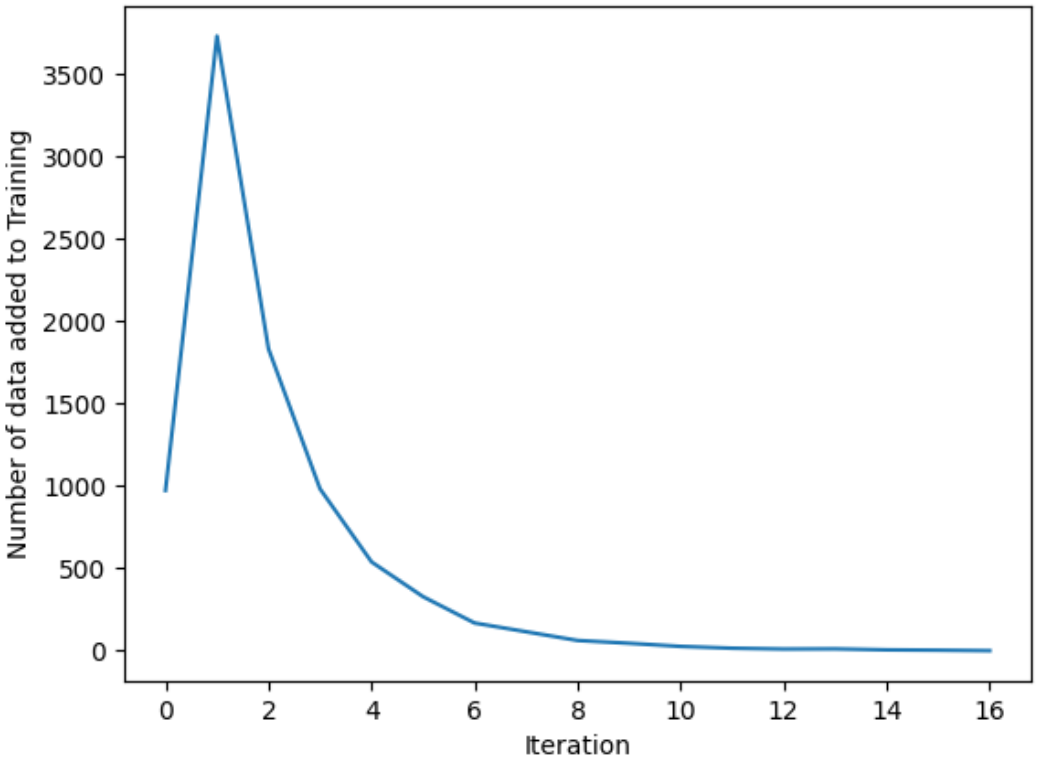
شکل ۳-۷- تعداد داده های برچسب زده شده در هر تکرار

شرط توقف الگوریتم برای قسمت های قبل وجود داشتن داده ی بدون برچسبی است که که در هر مرحله احتمال اختصاص داده شده به آن بیشتر از 99% باشد.

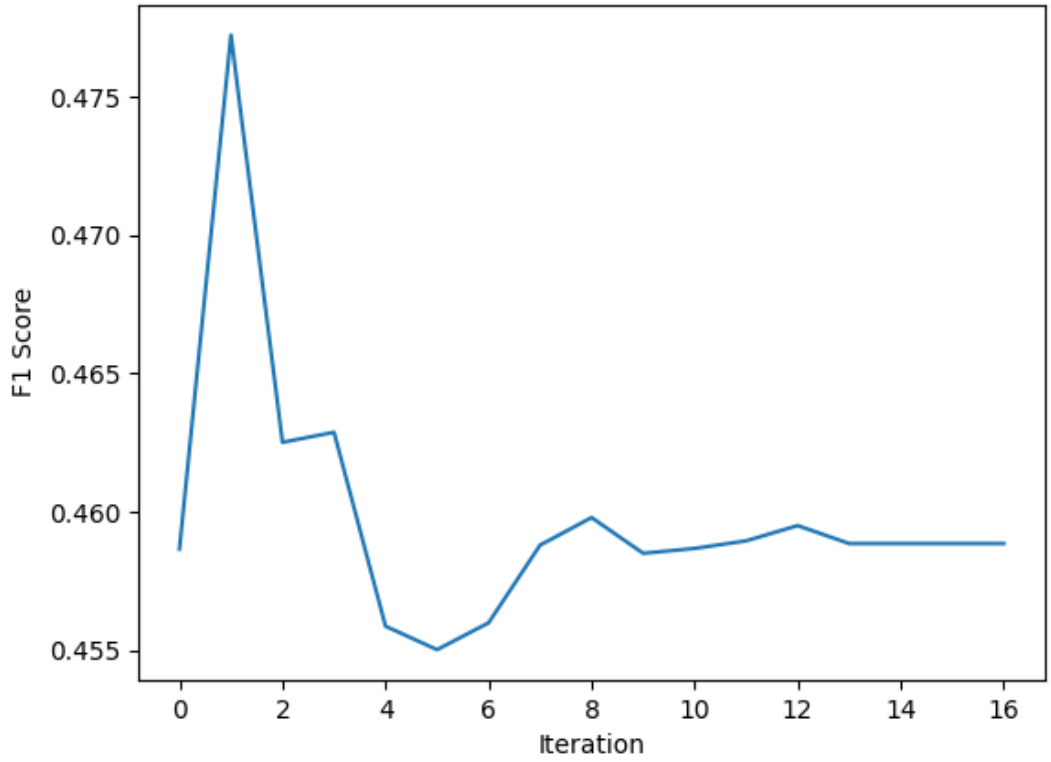
حال با عوض کردن این آستانه عملگرد الگوریتم خود را بررسی می کنیم:

```
Accuracy of the model is:  0.7537578573380705
F1 score of the model is:  [0.8406156  0.45885886]
```

شکل ۳-۸- دقت و معیار F1 به ازای آستانه ۹۵٪



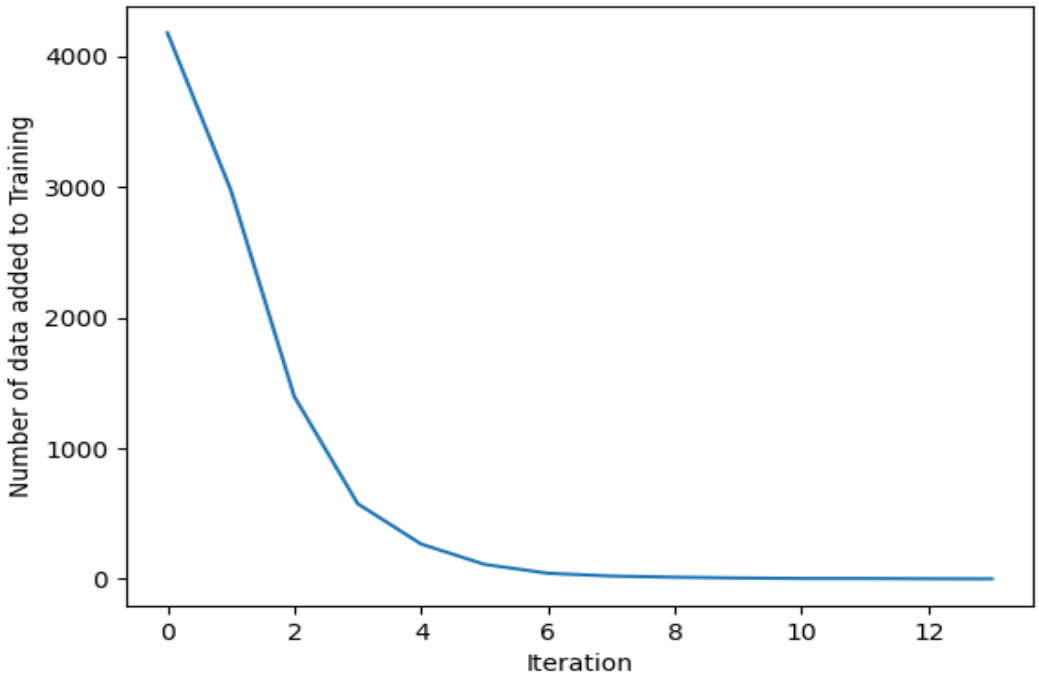
شکل ۳-۹- تعداد داده های برچسب زده شده در هر تکرار به ازای آستانه ۹۵٪



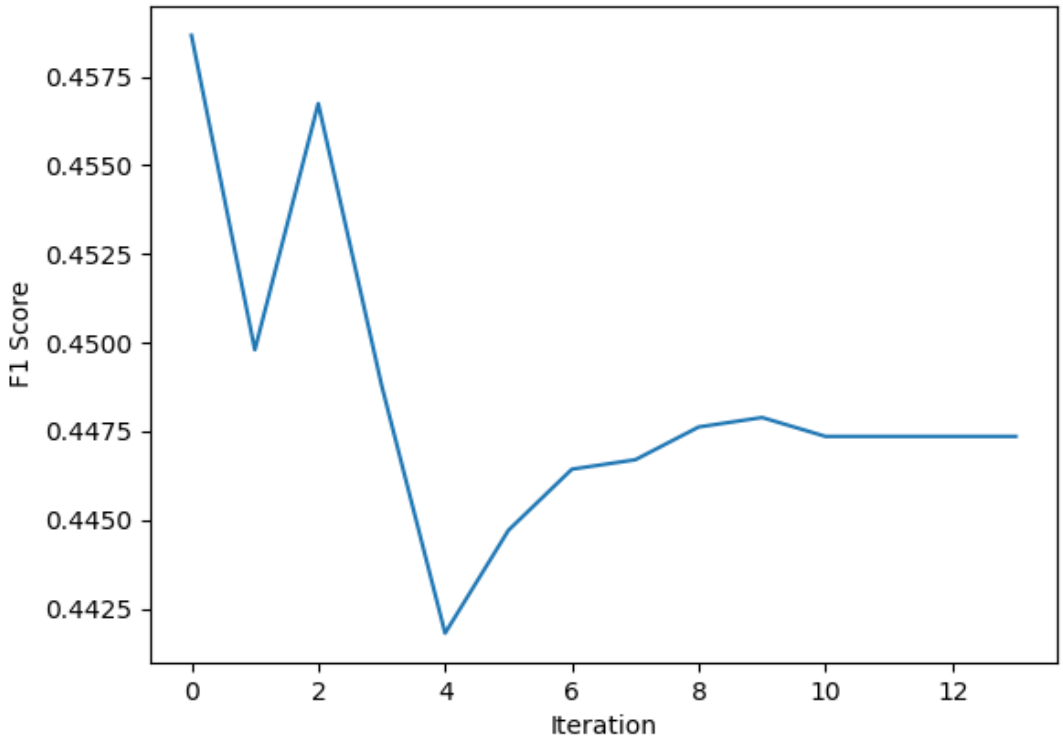
شکل ۳-۱۱- معیار F1 به ازای هر تکرار به ازای آستانه ۹۵٪

```
Accuracy of the model is:  0.7461054933041815
F1 score of the model is:  [0.83519603 0.44735277]
```

شکل ۳-۱۲- دقت و معیار F1 به ازای آستانه ۹۰٪



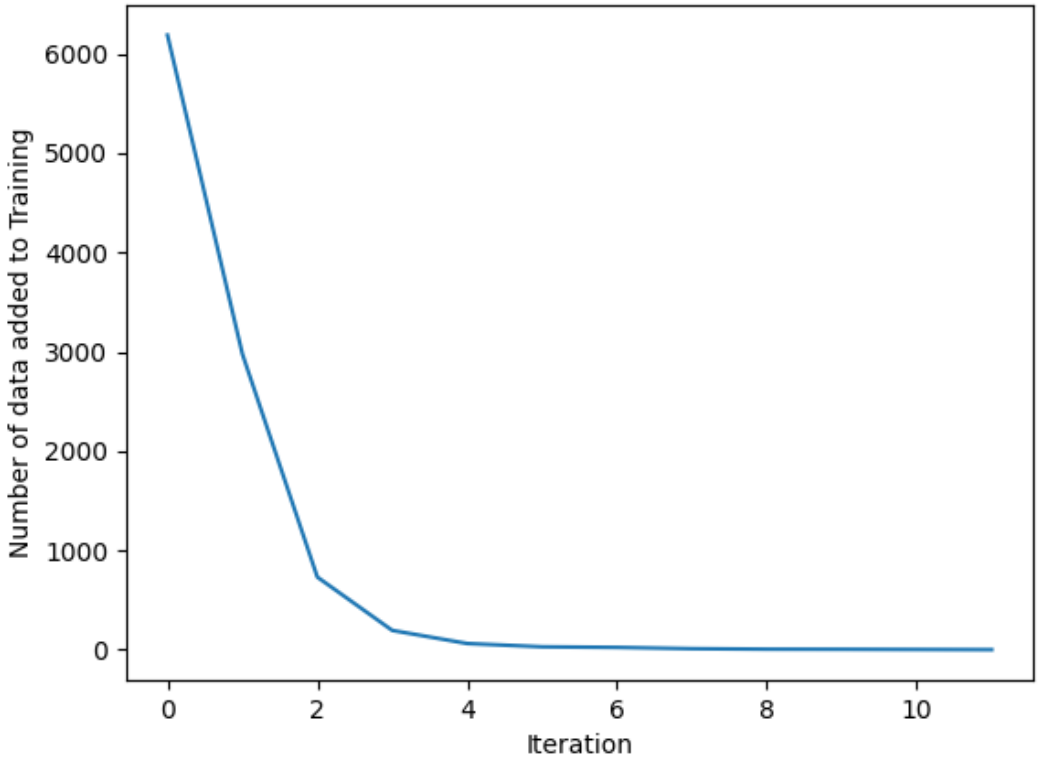
شکل ۳-۱۳- تعداد داده های برچسب زده شده در هر تکرار به ازای آستانه ۹۰٪



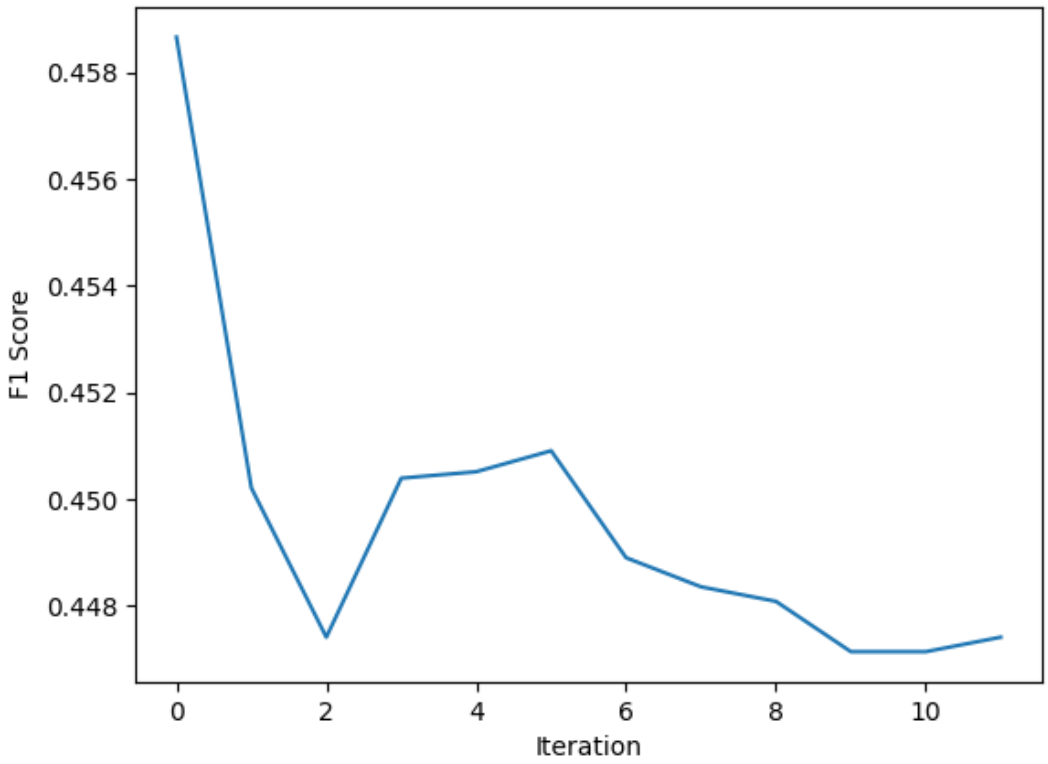
شکل ۳-۱۴- - معیار F1 به ازای هر تکرار به ازای آستانه ۹۰٪

```
Accuracy of the model is:  0.7512981688986061
F1 score of the model is:  [0.83956276 0.44714459]
```

شکل ۳-۱۵- دقت و معیار F1 به ازای آستانه ۸۰٪



شکل ۳-۱۶- تعداد داده های برچسب زده شده در هر تکرار به ازای آستانه ۸۰٪



شکل ۳-۱۷- - معیار F1 به ازای هر تکرار به ازای آستانه ۸۰٪

با توجه به شکل های بالا می توانیم نتیجه بگیریم که آستانه احتمال 0.99 بیشترین افزایش معیار F1 را داشته است پس این مقدار به نظر مناسب می آید.

سوال ۴

در قسمت اول این سوال به سوال های تحلیلی خواسته شده پاسخ می دهیم، سپس در قسمت ب-۱ مسئله تولد را به کمک پایتون شبیه سازی کرده و نمودار احتمال برابر بودن روز تولد به ازای تعداد افراد گروه را رسم می کنیم، در قسمت ب-۲ صحت قضیه حد مرکزی [1] را برای دو توزیع نمایی و دو جمله ای [2] به کمک شبیه سازی در پایتون بررسی می کنیم و موارد خواسته شده را رسم می کنیم.

الف-۱:

برای پرتاب دوبار سکه داریم:

- احتمال آمدن دوبار شیر: P^2
- احتمال آمدن یک بار شیر و یک بار خط: $2P(1-P)$
- احتمال آمدن دو بار خط: $(1-P)^2$

طبق قانون بیز داریم:

$$P\left(\text{دو سکه متفاوت} \middle| \text{پرتاب اول شیر}\right) = \frac{P\left(\text{پرتاب اول شیر} \mid \text{دو سکه متفاوت}\right).P(\text{شیر})}{P(\text{خط})}$$

$$= \frac{(1-P).P}{2P(1-P)} = \frac{1}{2}$$

$$P\left(\text{دو سکه متفاوت} \middle| \text{پرتاب اول خط}\right) = 1 - \frac{1}{2} = \frac{1}{2}$$

پس الگوریتم به این صورت شد که: سکه خراب را دوبار پرتاب می کنیم اگر نتیجه هر دو پرتاب متفاوت بود، نتیجه پرتاب اول را به عنوان خروجی در نظر می گیریم.

[1] Central Limit Theroem
[2] Binomial

الف-۲:

الف۲-۱:

$$f(x) = \lambda_x e^{-\lambda_x x} \quad for\ x > 0$$
$$f(y) = \lambda_y e^{-\lambda_y y} \quad for\ y > 0$$
$$E\{X\} = \frac{1}{\lambda_x}\ , E\{Y\} = \frac{1}{\lambda_y}$$
$$U = \min(X, Y)$$
$$V = \max(X, Y)$$

$$UV = XY$$

$$\rightarrow E\{UV\} = E\{XY\} = E\{X\}E\{Y\} = \frac{1}{\lambda_y \lambda_x}$$
$$P\{V\} = P\{X \leq v\ \&\ Y \leq v\} = F_x(v)F_y(v)$$
$$= \begin{cases} 0 & v \leq 0 \\ \left(1 - e^{-\lambda_x v}\right)\left(1 - e^{-\lambda_y v}\right) & v > 0 \end{cases}$$
$$\rightarrow f_v(v) = \begin{cases} 0 & z \leq 0 \\ \lambda_x e^{-\lambda_x v} + \lambda_y e^{-\lambda_y v} - (\lambda_x + \lambda_y) e^{-(\lambda_y + \lambda_x)v} \end{cases}$$
$$\rightarrow E\{V\} = \int v f_v(v)\, dv = \frac{\lambda_x^2 + \lambda_x^2 + \lambda_y \lambda_x}{\lambda_y \lambda_x (\lambda_y + \lambda_x)}$$
$$E\{V + U\} = E\{X + Y\} = \frac{1}{\lambda_x} + \frac{1}{\lambda_y} \rightarrow E\{U\} = \frac{1}{\lambda_x} + \frac{1}{\lambda_y} - \frac{\lambda_x^2 + \lambda_x^2 + \lambda_y \lambda_x}{\lambda_y \lambda_x (\lambda_y + \lambda_x)} = \frac{1}{(\lambda_y + \lambda_x)}$$
$$\rightarrow cov(U, V) = E\{UV\} - E\{U\}E\{V\} = \frac{1}{\lambda_y \lambda_x} - \frac{\lambda_x^2 + \lambda_x^2 + \lambda_y \lambda_x}{\lambda_y \lambda_x (\lambda_y + \lambda_x)} \cdot \frac{1}{(\lambda_y + \lambda_x)} = \frac{1}{(\lambda_y + \lambda_x)^2}$$

الف۲-۲:

$$U \sim N(\mu, \sigma^2), V \sim N(\mu, \sigma^2)$$

$$Z = \max(U, V)$$

$$E[min(X,Y)] = \mu - \frac{\sigma}{\sqrt{\pi}}, E[max(X,Y)] = \mu + \frac{\sigma}{\sqrt{\pi}}$$

$$cov(V, \max(U,V) = cov(V,Z) = E\{VZ\} - E\{V\}E\{Z\}$$

$$E\{VZ\} = E\{V^2 | V > U\} + E\{VU\,|U > V\}$$

$$E\{VU\,|U > V\} + E\{VU\,|U < V\} = E\{UV\} = \mu^2 \rightarrow E\{VU\,|U > V\} = \frac{\mu^2}{2}$$

$$\rightarrow E\{VZ\} = E\{V^2F(V)\} + \frac{\mu^2}{2} = E\{(-V)^2F(-V)\} + \frac{\mu^2}{2} = E\{V^2\} - E\{V^2F(V)\} + \frac{\mu^2}{2}$$

$$\rightarrow cov(V, \max(U,V)) = \frac{1}{2}var(V) + \frac{\mu^2}{2} - \mu\left(\mu + \frac{\sigma}{\sqrt{\pi}}\right)$$

به طور مشابه:

$$cov(V, \min(U,V)) = \frac{1}{2}var(V) + \frac{\mu^2}{2} - \mu\left(\mu - \frac{\sigma}{\sqrt{\pi}}\right)$$

اگر توزیع نرمال ما استاندارد باشد روابط بالا به صورت زیر ساده می شوند:

$$cov(V, \max(U,V)) = \frac{1}{2}var(V) = 0.5$$

$$cov(V, \min(U,V)) = \frac{1}{2}var(V) = 0.5$$

الف-۳:

ابتدا c را پیدا می کنیم:

$$\iint f_{xy}(x,y)dxdy = 1$$

$$c\int_0^1 \frac{(x-1)^2}{2} = 1 \rightarrow c = 6$$

$$f_x = \int_0^{1-x} 6(1-y-x)dy = 3(x-1)^2$$

$$f_y = \int_0^{1-y} 6(1-y-x)dx = 3(y-1)^2$$

$$\Pr\{X \le 0.5\} = \int_0^{0.5} 3(x-1)^2 dx = \frac{7}{8}$$

$$E\{X+Y\} = E\{X\} + E\{Y\} = \int_0^1 3x(x-1)^2 + \int_0^1 3y(y-1)^2 = \frac{1}{2}$$

$\Pr\{X \le 0.5\}$ احتمال کار کردن X کمتر یا مساوی 0.5 (یعنی احتمال اینکه X کمتر یا مساوی نصف کار را انجام دهد) را بیان می کند

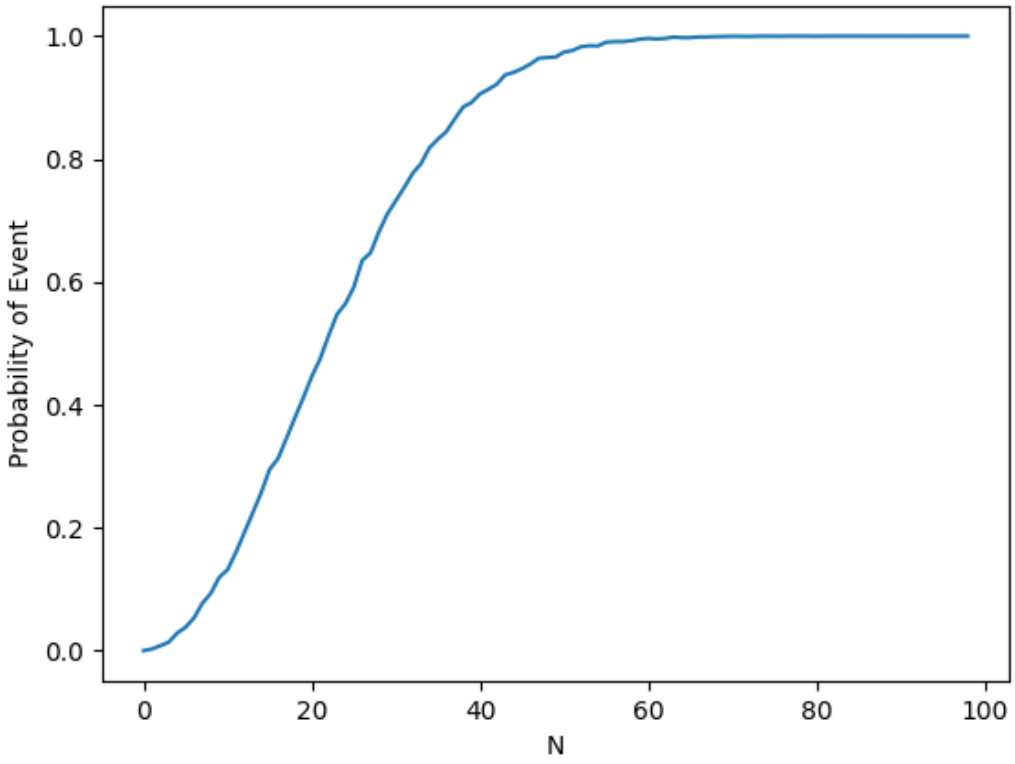
$E\{X+Y\}$ میانگین کار انجام شده توسط دو کارمند.

ب-۱:

عدد n را به طور **تصادفی** 69 در نظر می گیریم، و احتمال خواسته شده را به دست می آوریم.

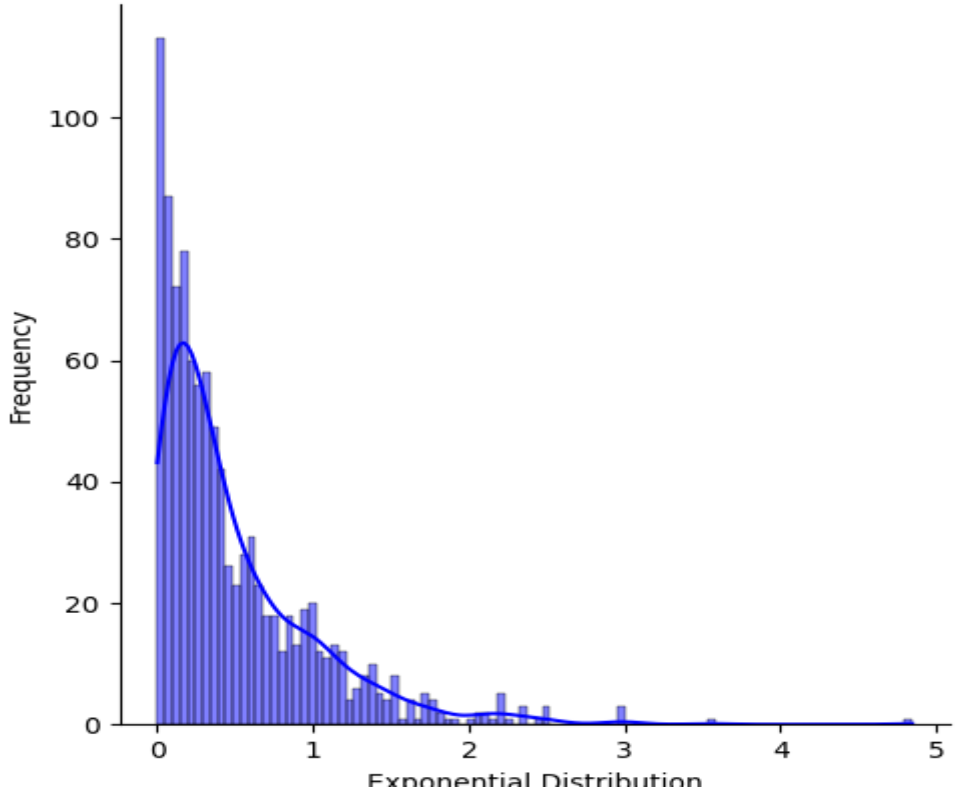
```
Probability that at least two people have same birthday in a group of 69 is:  0.9985
```

شکل ۴-۱- احتمال برابر بودن روز تولد حداقل دو نفر در یک گروه ۶۹ نفره

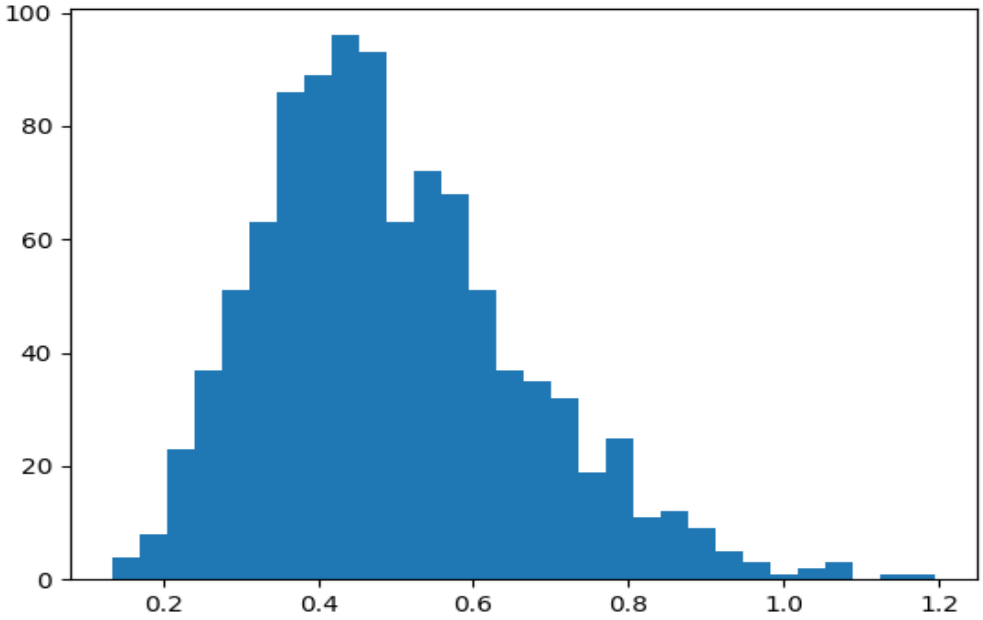


شکل ۴-۲- نمودار احتمال برابر بودن روز تولد حداقل دو نفر بر اساس تعداد آنها

ب-۲:



شکل ۴-۳- نمودار توزیع نمایی با نرخ ۲



شکل ۴-۴- نمودار توزیع میانگین داده های نمونه برداری شده از توزیع نمایی با نرخ ۲

با توجه به شکل۴-۴ نمودار توزیع میانگین نمونه های برداشته شده بسیار شبیه توزیع نرمال با میانگین حدود 0.5 است.

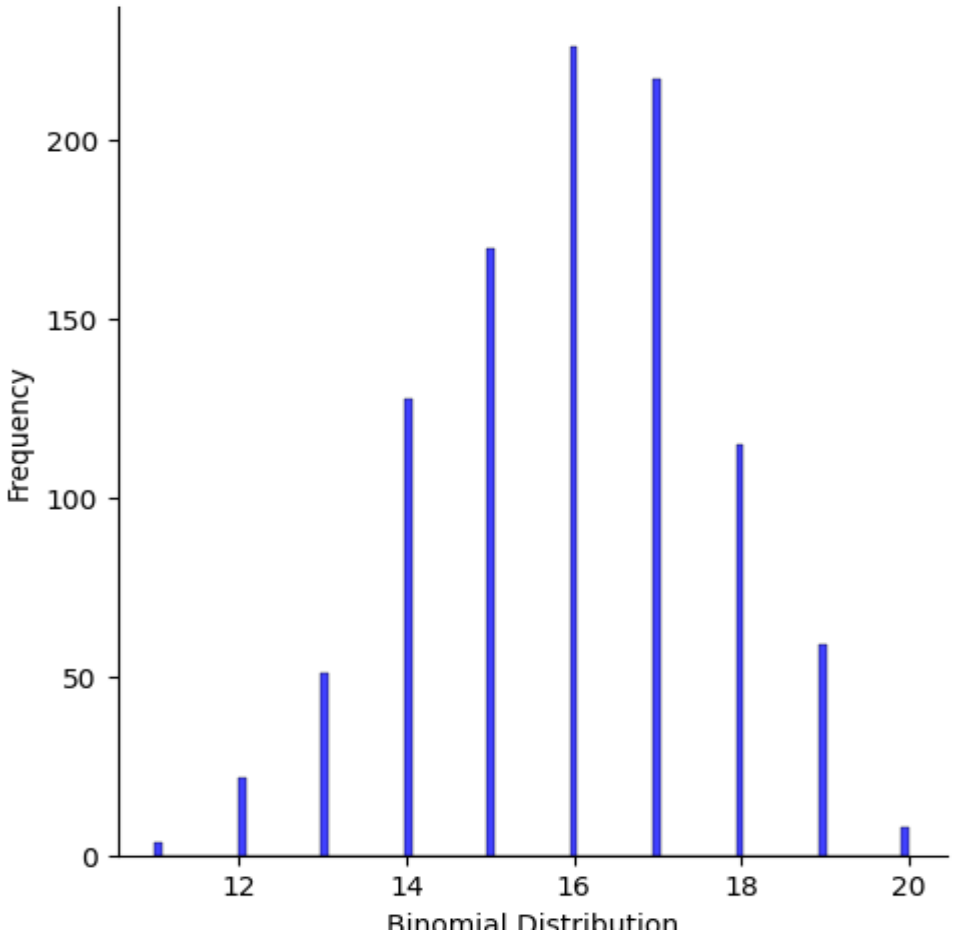
```
Mean of the samples is equal to:  0.500907029353346
Standard deviation of the samples is equal to:  0.5002930897358169
```

شکل ۴-۵- دقت و انحراف معیار داده های نمونه برداری شده از توزیع نمایی با نرخ ۲

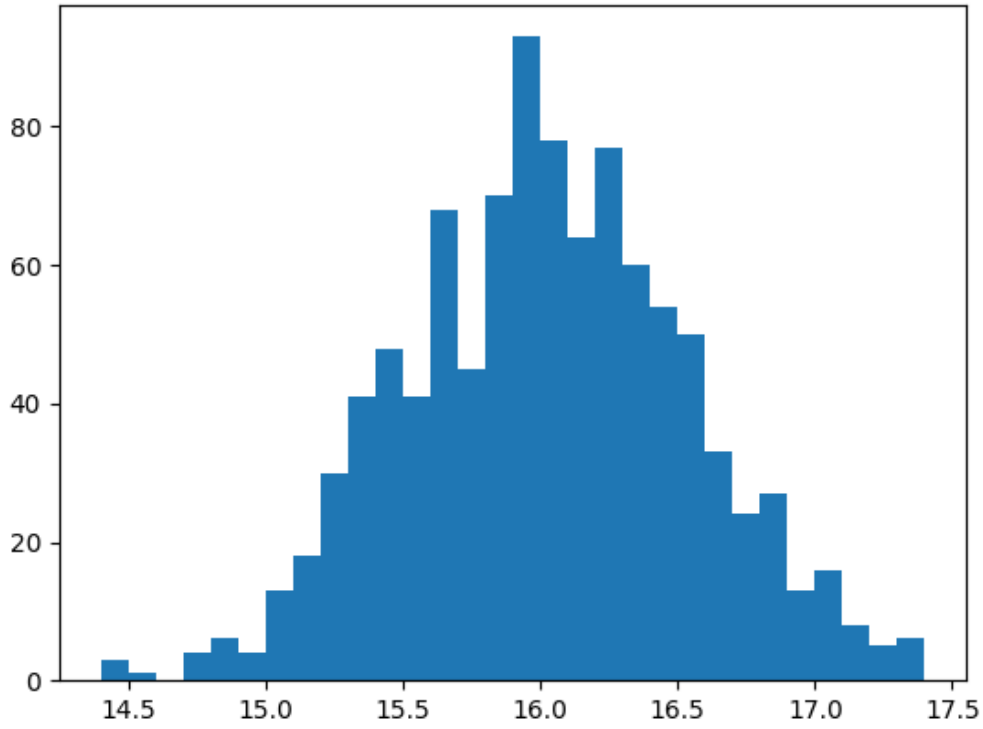
همچنین می دانیم در توزیع نمایی با نرخ ۲ میانگین و انحراف معیار به صورت زیر است:

$$\mu = \sigma = \frac{1}{\lambda} = \frac{1}{2}$$

که با مقدار میانگین و انحراف معیاری که از قضیه حد مرکزی به دست می آید مطابقت دارد.



شکل ۴-۶- نمودار توزیع دو جمله ای



شکل ۴-۷- نمودار توزیع میانگین داده های نمونه برداری شده از توزیع دو جمله ای

با توجه به شکل۴-۷ نمودار توزیع میانگین نمونه های برداشته شده بسیار شبیه توزیع نرمال با میانگین حدود 16 است.

```
Mean of the samples is equal to:  16.1039
Standard deviation of the samples is equal to:  1.7436469797524956
```

شکل ۴-۸- دقت و انحراف معیار داده های نمونه برداری شده از توزیع دو جمله ای

همچنین می دانیم در توزیع دوجمله ای با نرخ n=20, p=0.8 میانگین و انحراف معیار به صورت زیر است:

$$\mu = np = 16, \qquad \sigma = \sqrt{np(1-p)} = \sqrt{3.2} = 1.78$$

که با مقدار میانگین و انحراف معیاری که از قضیه حد مرکزی به دست می آید مطابقت دارد.

پیوست

کد الگوریتم خوشه بندی سلسله مراتبی [1] و رسم نمودار درختی برای سوال ۱ قسمت ب

```
import pandas as pd
import matplotlib.pyplot as plt
import scipy.cluster.hierarchy as shc
from scipy.spatial.distance import squareform, pdist

point = ['P1','P2','P3','P4','P5']
data = pd.DataFrame({'Point':point, 'a':[0.22,0.35,0.26,0.08,0.45],
'b':[0.38,0.32,0.19,0.41,0.3]})
data = data.set_index('Point')
print(data)
plt.scatter(data['a'], data['b'], c='r', marker='*')
for j in data.itertuples():
    plt.annotate(j.Index, (j.a, j.b), fontsize=15)
plt.show()
dist = pd.DataFrame(squareform(pdist(data[['a', 'b']]), 'euclidean'),
columns=data.index.values, index=data.index.values)
print(dist)
plt.figure(figsize=(12,5))
plt.title("Dendrogram with Single linkage")
dend = shc.dendrogram(shc.linkage(data[['a', 'b']], method='single'),
labels=data.index)
plt.show()
```

منبع:

https://www.analyticsvidhya.com/blog/2021/06/single-link-hierarchical-clustering-clearly-explained/

[1] Hierarchical Clustering