دانشگاه تهران

پردیس دانشکده های فنی

دانشکده مهندسی برق و کامپیوتر

یادگیری ماشین

تمرین سوم

امیرحسین عباسکوهی

۸۱۰۱۹۷۵۳۹

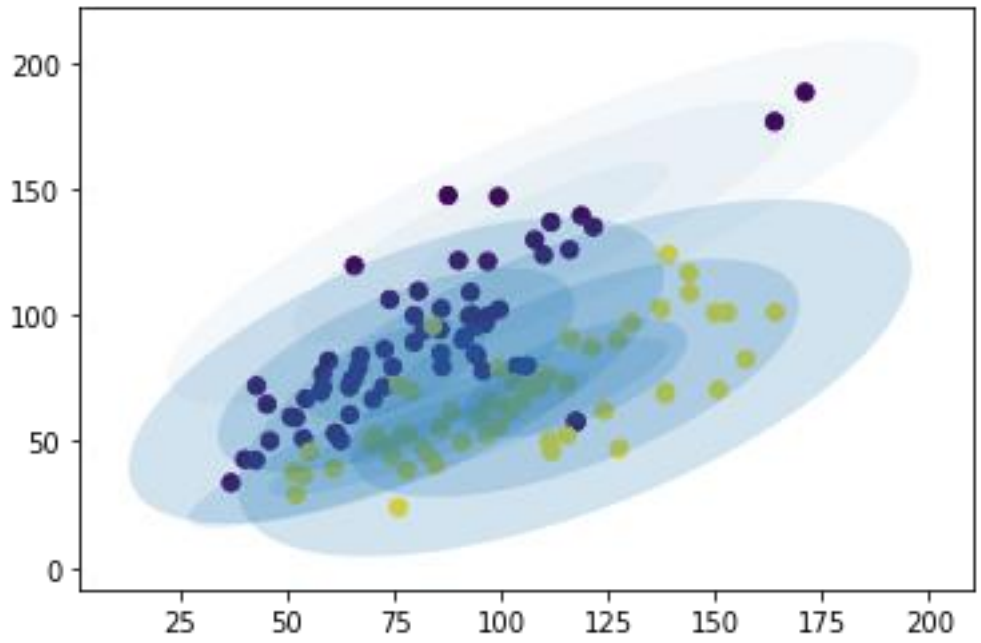
استاد ابولقاسمی

سوال ۸)

برای این بخش از تمرین باید از تصاویر دو تیم منچستر و چلسی برای ایجاد یک طبقه بند استفاده کنیم. بر اساس اطلاعاتی که از تمرین یک که در آن جا هم از همین داده ها استفاده شده بود باز هم برای هر عکس دو ویژگی تعریف میکنیم: ویژگی کانال قرمز و ویژگی کانال آبی که این ویژگی ها میانگین این دو کانال برای تمام پیکسل های داده است. برای این منظور تابع rgb_mean استفاده شده است.

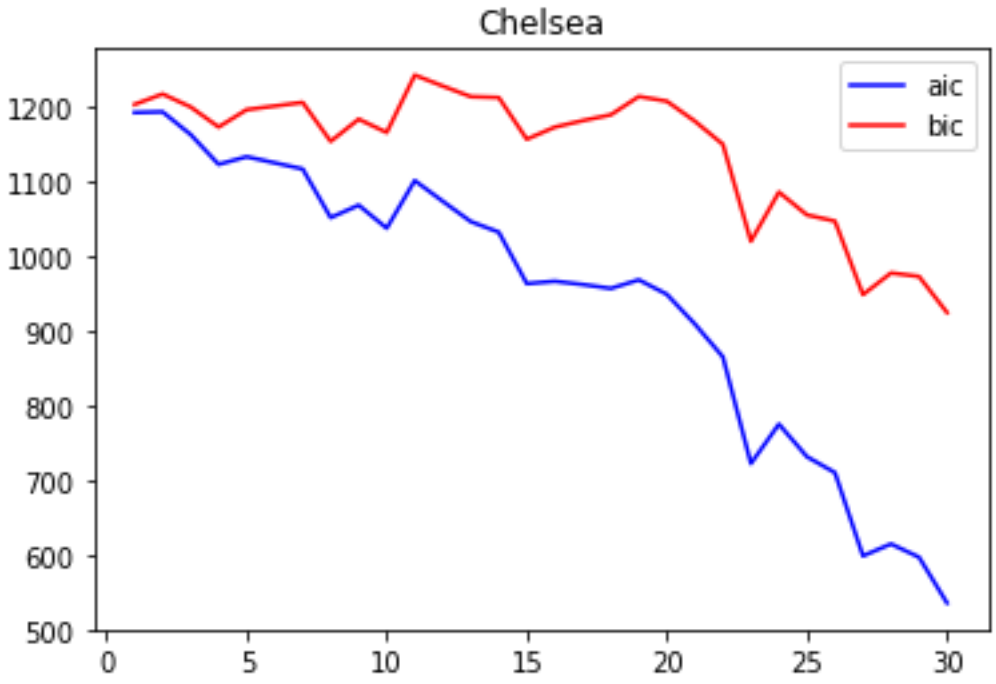
سپس دیتاستی که ساختیم را برای دو کلاس مجزا میکنیم تا تمرین مدل ها را جدا انجام بدهیم برای هر کدام.

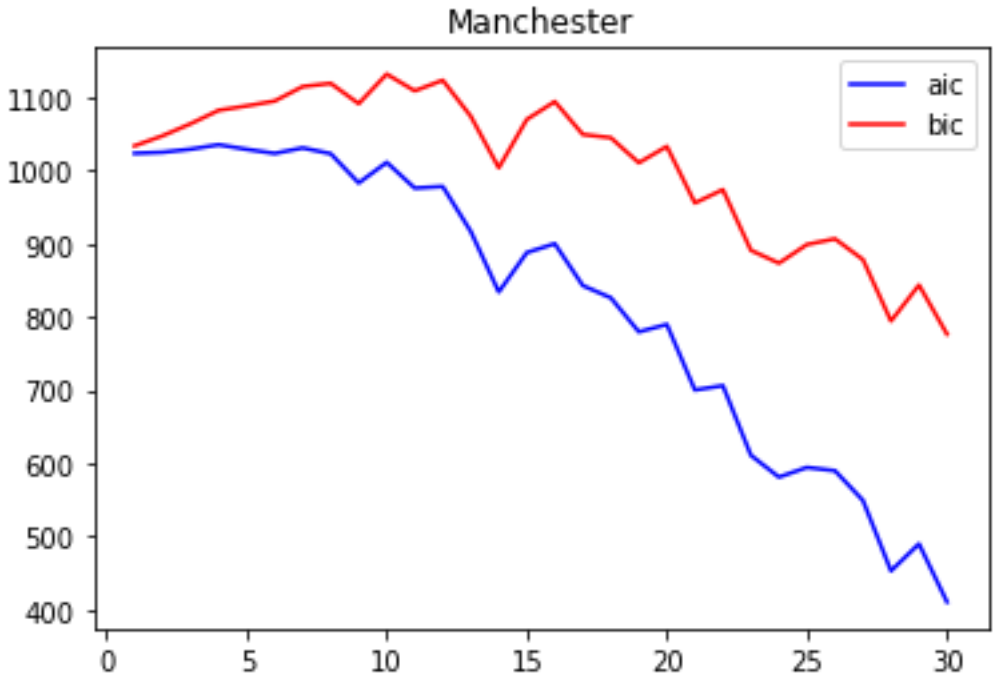
برای مدل از کتابخانه sklearn بخش mixture و از GaussianMixture استفاده میکنیم. برای هر کدام از دو کلاس دو مدل GMM با دو کامپوننت تمرین میدهیم ونتایج را با استفاده از matplotlib رسم میکنیم. خروجی را در اینجا میبینیم:



در نهایت باید برای پارامتر های مختلف aic و bic ها را به دست اوریم. به همین دلیل تعداد کامپوننت ها را از ۱ تا ۳۰ مقدار دهی میکنیم و در هر مرحله بعد از تمرین داده کلاس از متد

aic و bic استفاده میکنیم و لیست را پرمیکنیم و در نهایت نمودار ها را میکشیم که به فرم زیر می شود:





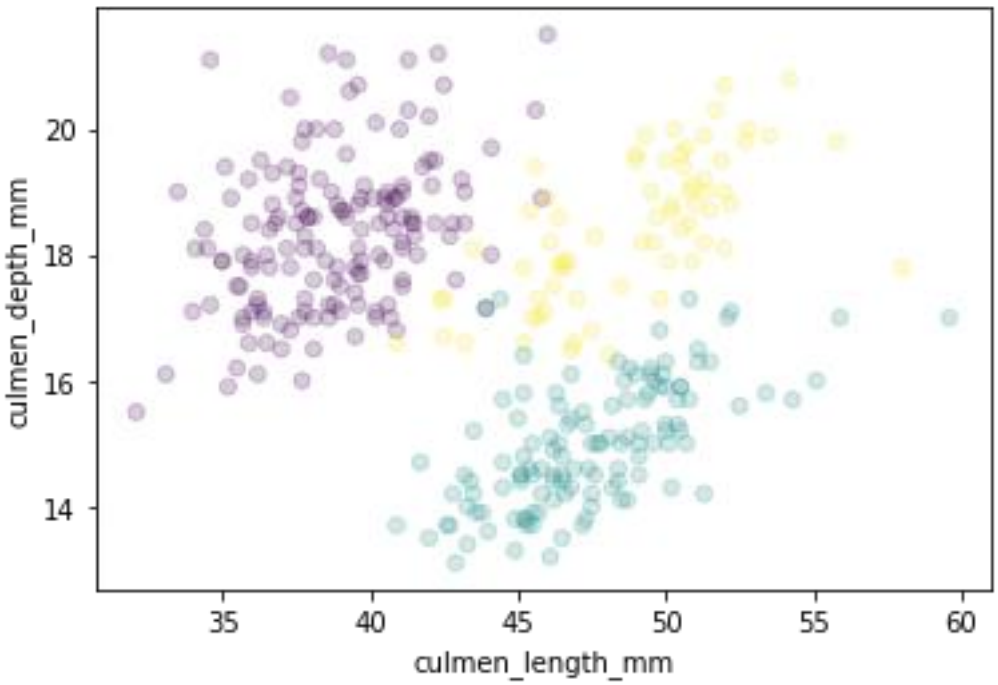
همانطور که میدانیم هر چه قدر مقدار bic وaic کمتر باشد نتیجه بهتر است که با بررسی انجام شده در بین این ۳۰ مقدار بهترین حالت ۲۷ است(البته برای مقادیر بیشتر همن تست شد و این

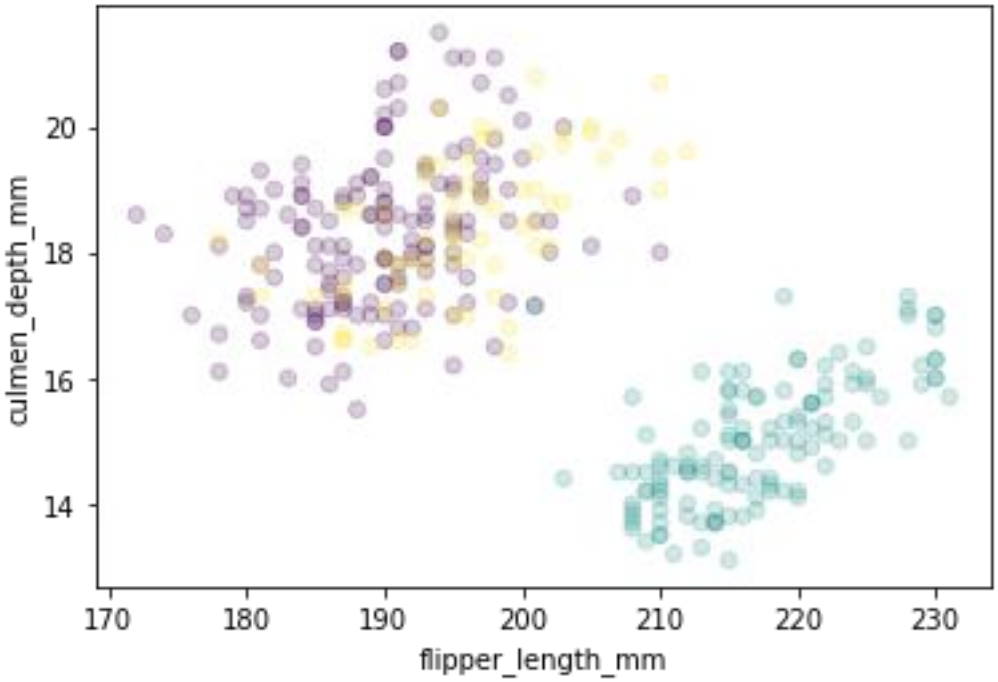
مقدار به طور نزولی کاهش پیدا میکند یعنی به طور کلی با افزایش تعداد کامپوننت اینجا نتیجه بهتر می شود. البته اگر برای کل داده اینکار را انجام میدادیم قطعا دو کامپوننت بهتر بود چون ۲ کلاس داریم. اینکه برای هر کدام جدا ترین را انجام داده ایم به خاطر پاسخ شما در تلگرام بوده است.)

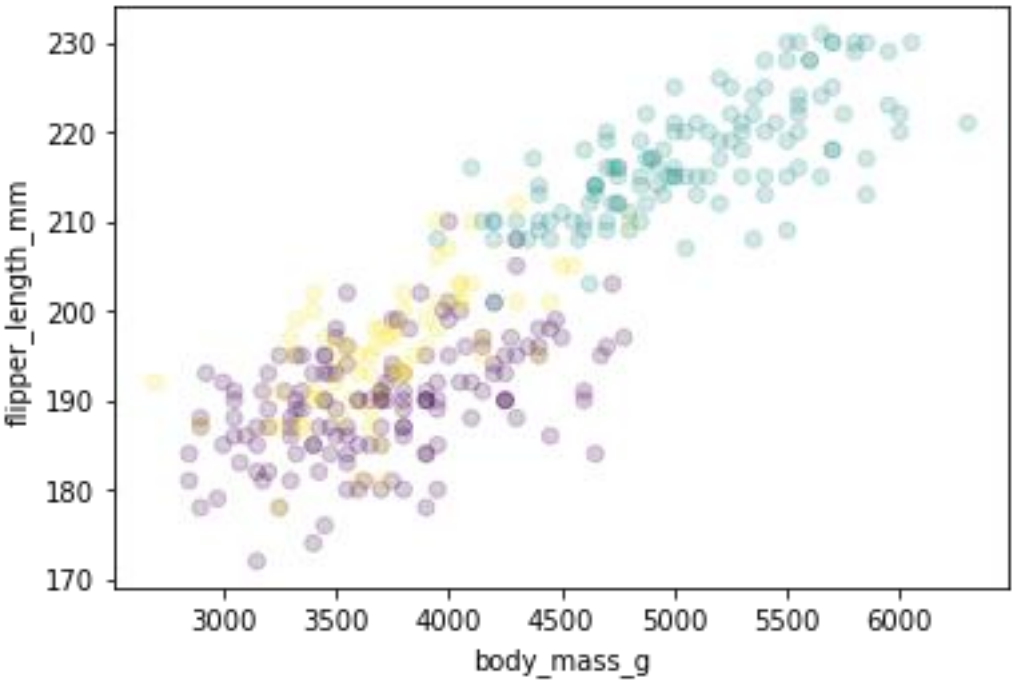
سوال ۹)

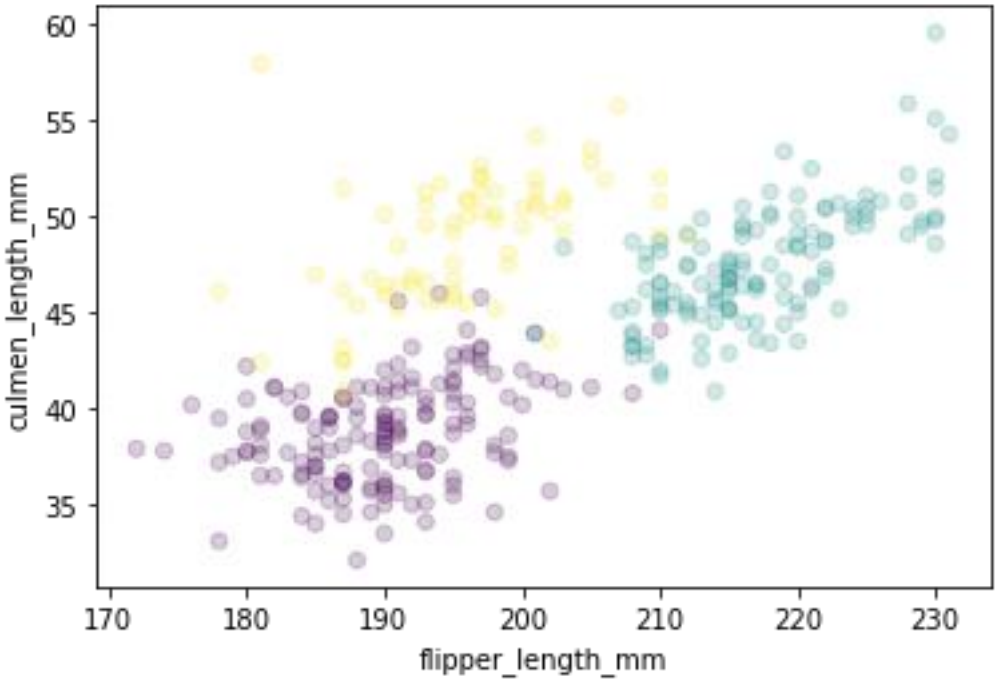
در این سوال یک دیتاست سه کلاسه داریم. البته در این دیتاست ما تعدادی NaN داریم و از طرفی هم با داده های categorical مواجه هستیم. به همین دلیل کاری که میکنیم این است که این موارد را در ابتدا هندل میکنیم.

در نهایت با استفاده از scatter از matplotlib داده ها را بر اساس دو ویژگی هایی که صورت سوال گفته شده است نمایش می دهیم که نتیجه در زیر آمده است:

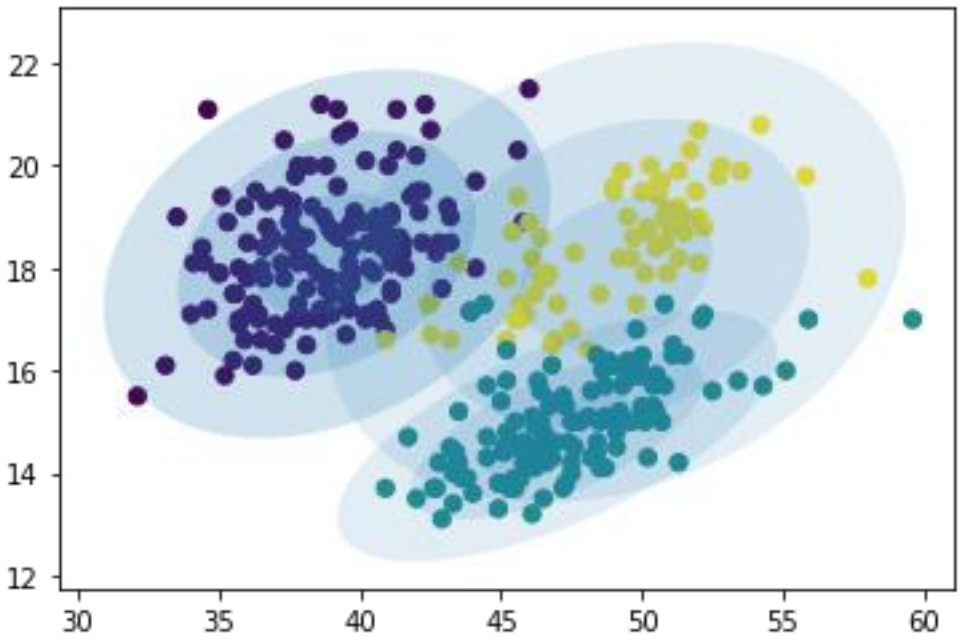



culmen_depth_mm
20
18
16
14
170
180
190
200
210
220
230
flipper_length_mm

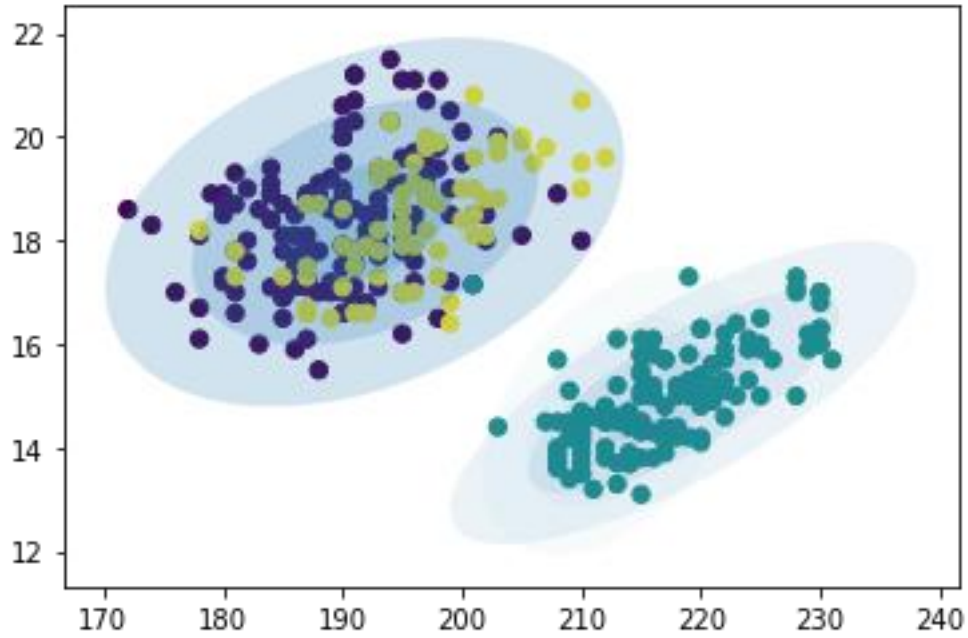

flipper_length_mm
230
220
210
200
190
180
170
3000
3500
4000
4500
5000
5500
6000
body_mass_g

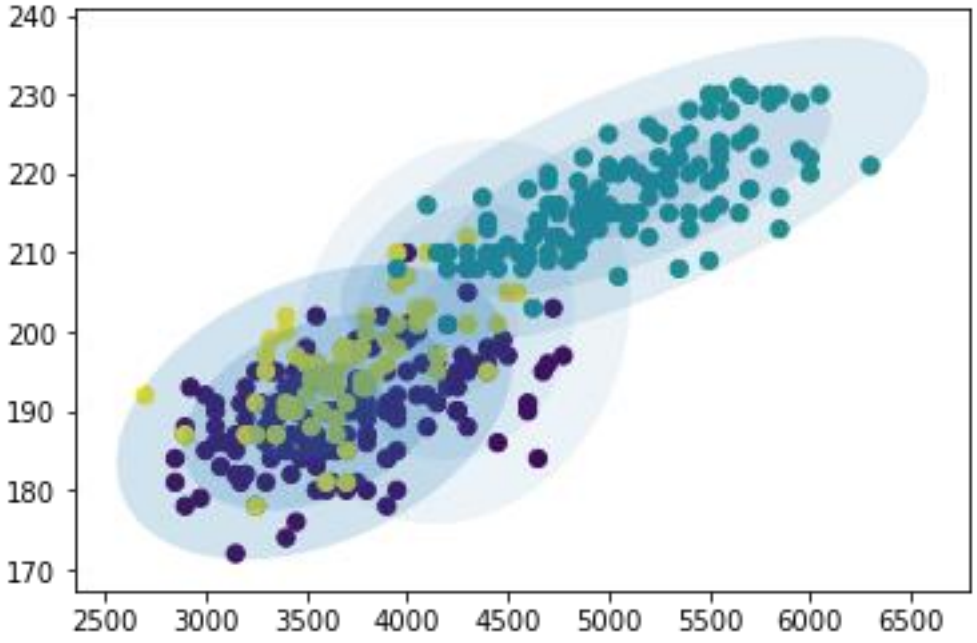


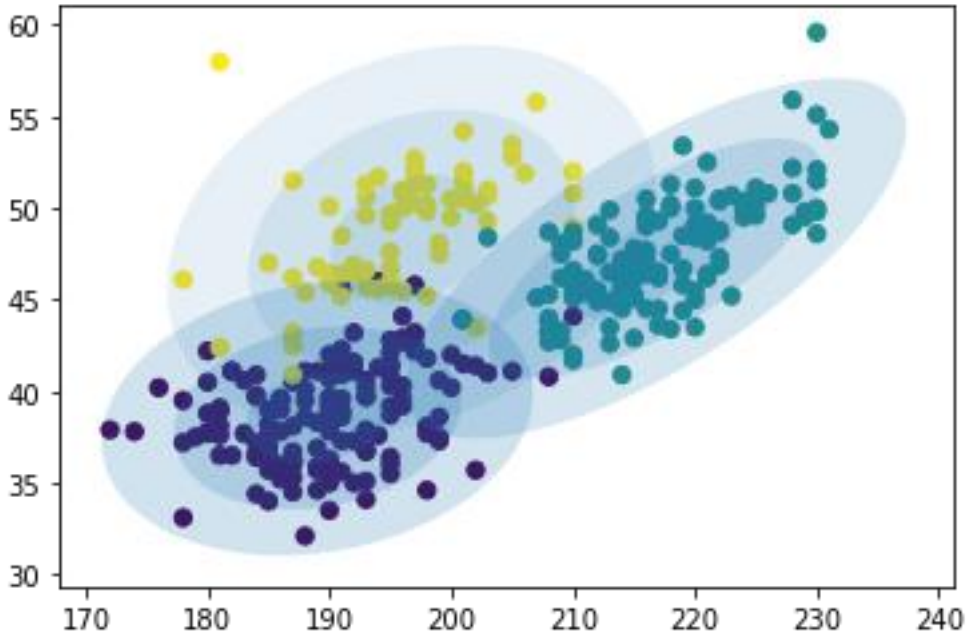
حال باید مدل GMM را ترین کنیم. کل داده ها را به دلیل داشتن سه کلاس با n_componnent=3 تمرین میدهیم و در نهایت پس از ترین شدن مدل با استفاده از فیچر های مدل یعنی mean،cov و weights کانتور ها را رسم میکنیم که نتایج در زیر آمده است:



v


22
20
18
16
14
12
170
180
190
200
210
220
230
240


240
230
220
210
200
190
180
170
2500
3000
3500
4000
4500
5000
5500
6000
6500

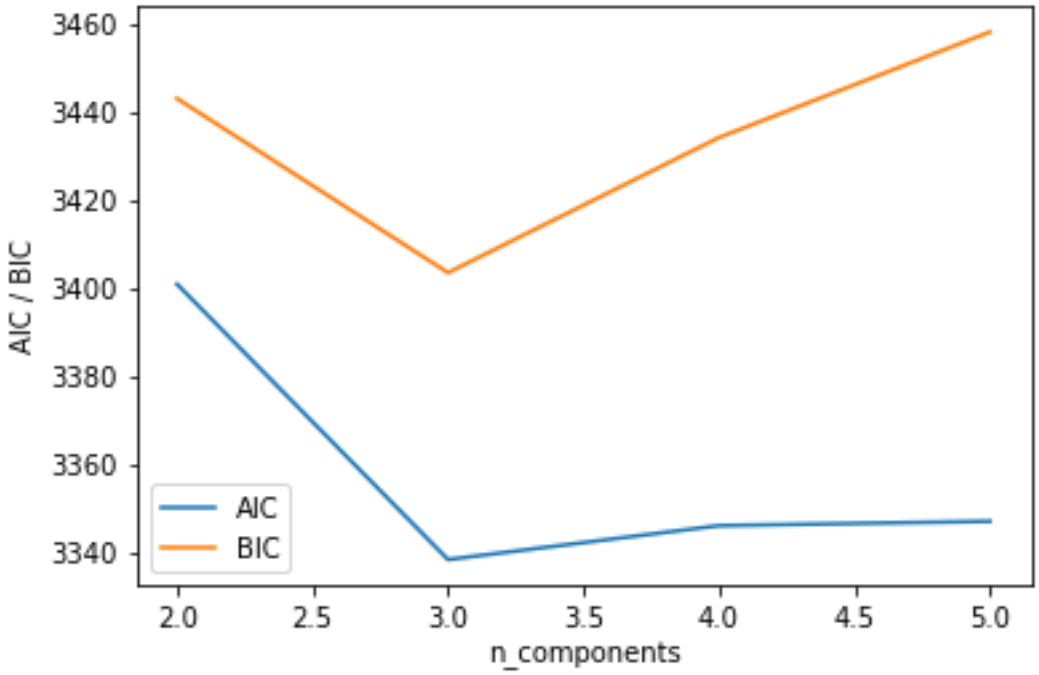


که در اینجا میبینیم که برای دو دسته از دو ویژگی ها ما در هم رفتگی داریم و مدل خوب کار نمیکند اما برای دو دسته دیگر مشکلی نداریم.

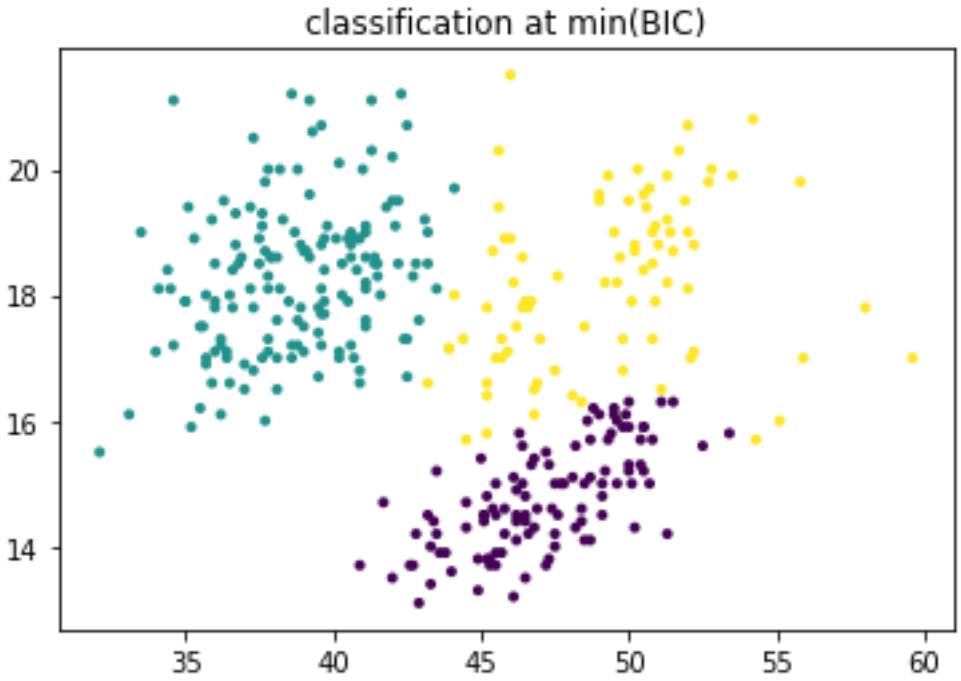
حال با استفاده از دستور predict کلاس بندی را انجام میدهیم و accuracy را حساب میکنیم. در صورت پروژه خطا گفته شده اما میدانیم که خطا در واقع همان 1-accuracy می باشد.

با بررسی های انجام شده بهترین ترکیب، flipper_length_mm and culmen_length_mm می باشد.

در نهایت هم مانند سوال قبل برای k های گفته شده در سوال مدل را تمرین می دهیم و aic/bic را با استفاده از متد های aic و bic در مدل GMM رسم میکنیم که نتیجه به صورت زیر می شود که نشان دهنده این است که k=3 بهترین است(چون ۳ کلاس داریم)

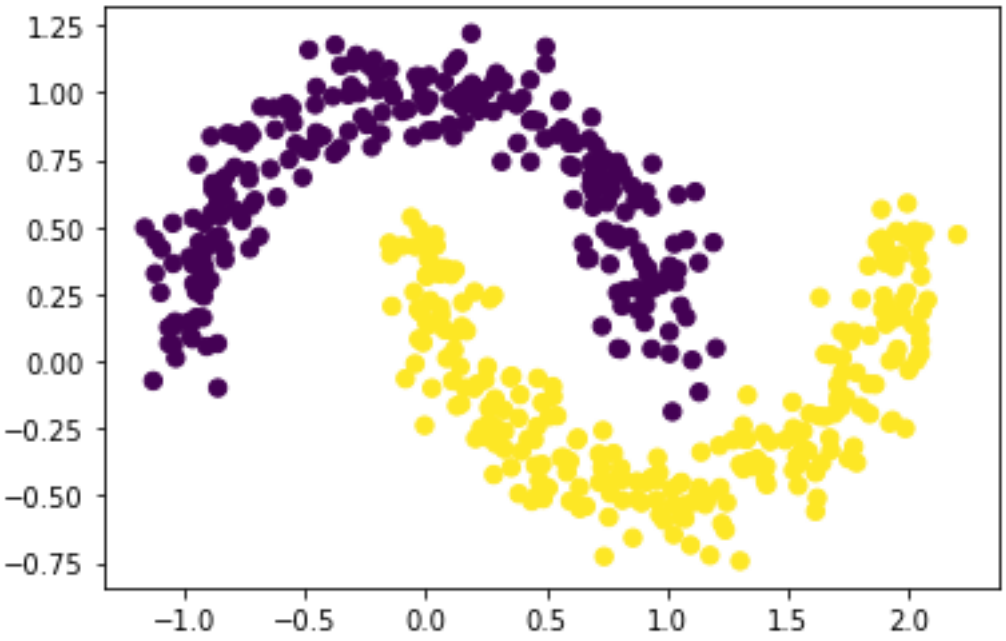


ودر زیر هم نتیجه این بهترین طبقه بند را میبنیم(نقاط پیش بینی شده هستند و نه نقاط در دیتاست):

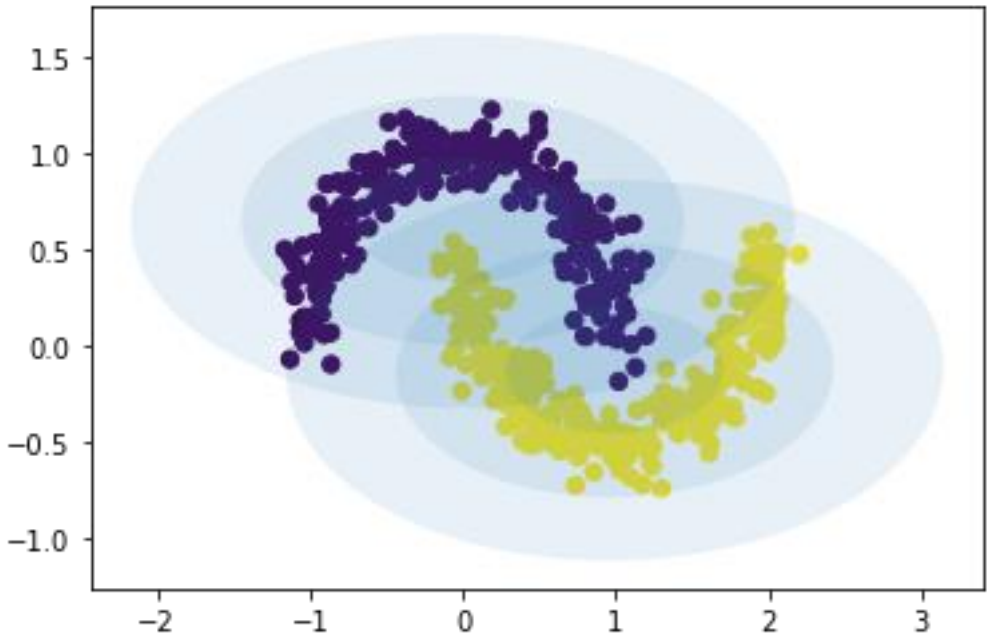


سوال ۱۰)

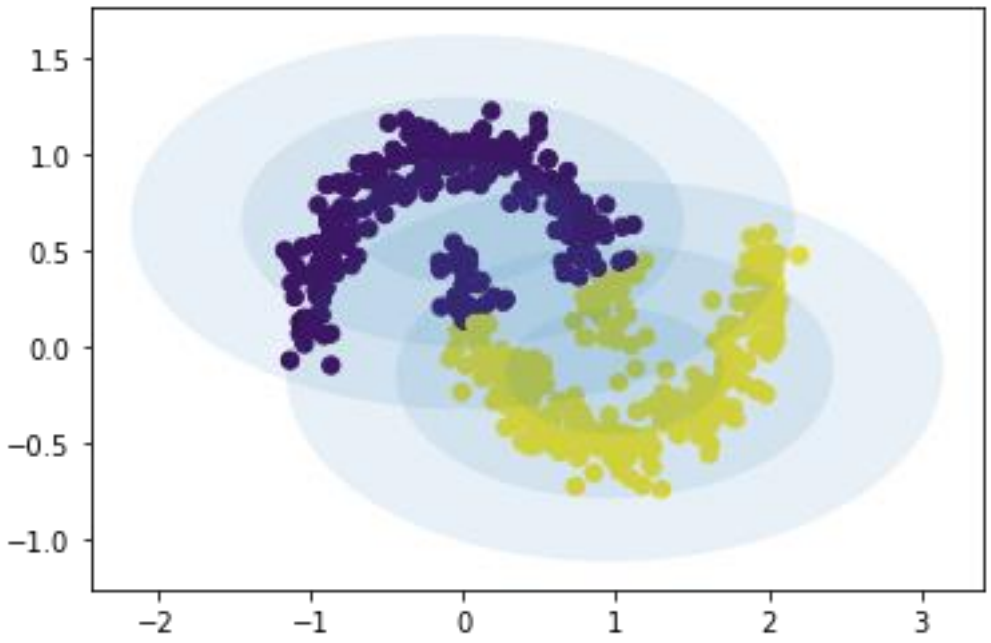
درابیتدا ۵۰۰ داده از دیتاست moon انتخاب می کنیم که داده ها به صورت زیر می باشد:



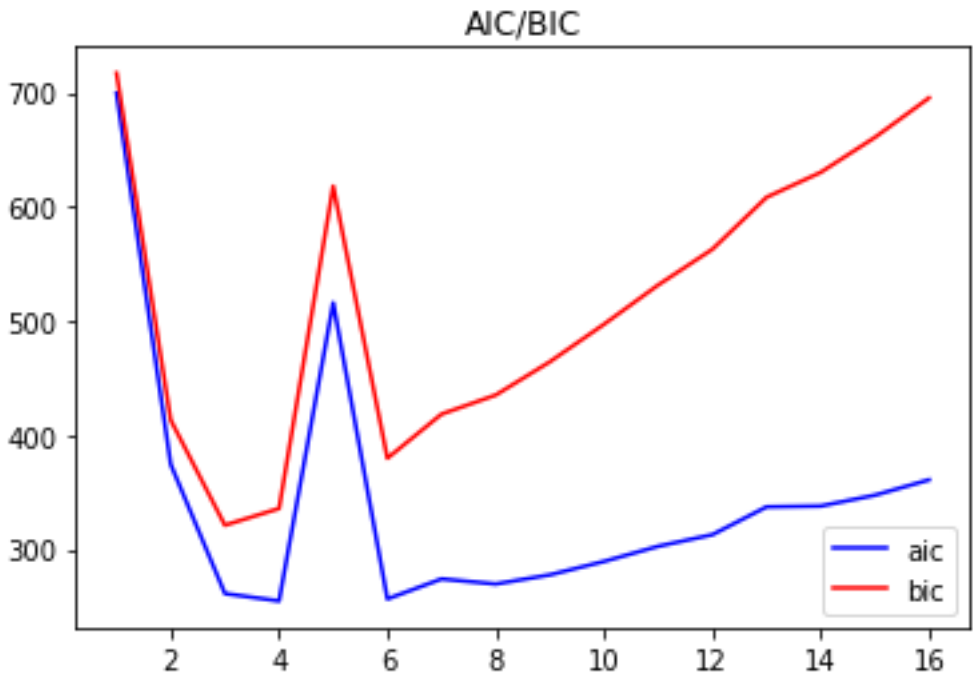
حال در اینجا با استفاده از رابطه گفته شده در کلاس از bayes estimation استفاده می کنیم و یک توزیع گوسی برای هر کلاس ایجاد می کنیم. البته باید دقت شود در اینجا ما حالت multivariant داریم. بدون کلاس بندی شکل داده ها به صورت زیر می باشد:

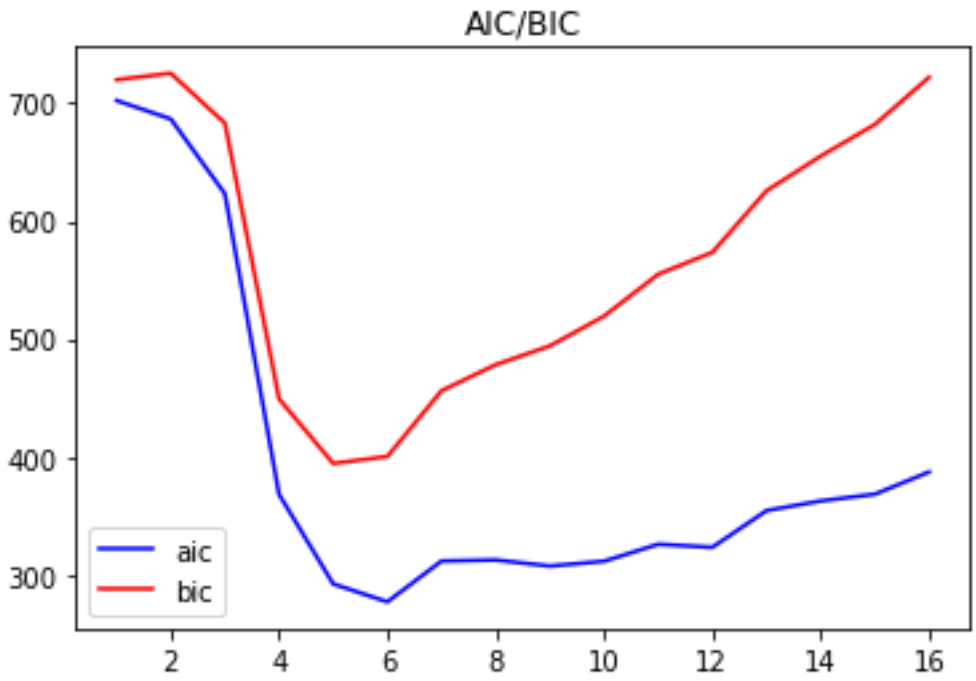


اما اگر pdf هر یک از توزیع ها را حساب کنیم و بر اساس بررسی احتمال کلاسی بندی کنیم نمودار زیر را داریم که کاملا منطقی است:



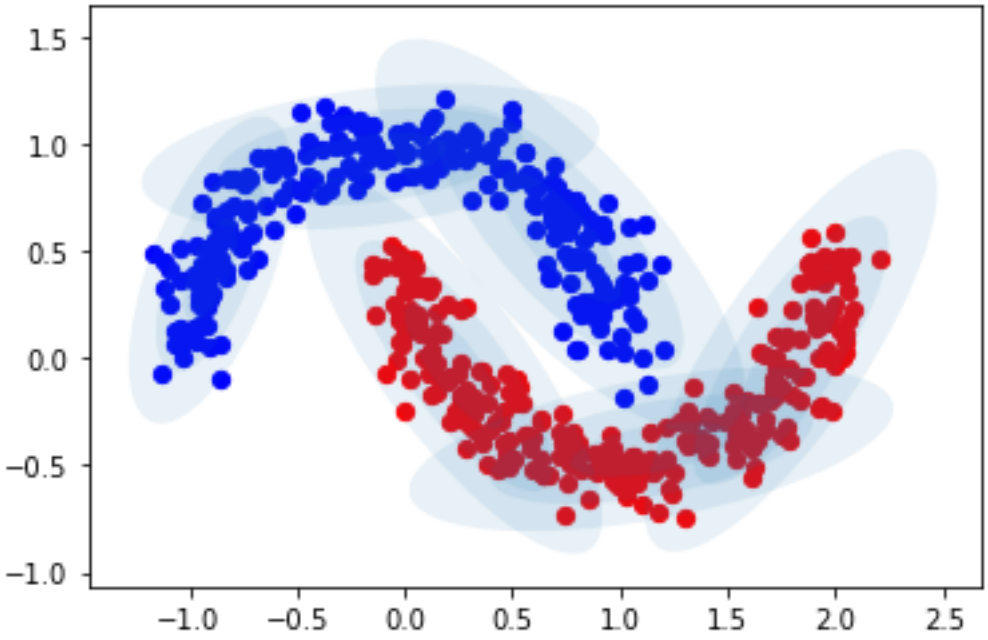
برای بخش بعد هم کلاس GMM را که از EM استفاده می کند پیاده سازی میکنیم و البته aic و bic را هم قرار میدهیم. حال این کلاس را برای پارامتر های مختلف حساب میکنیم و نمودار aic و bic برای دو کلاس به صورت زیر می شود:

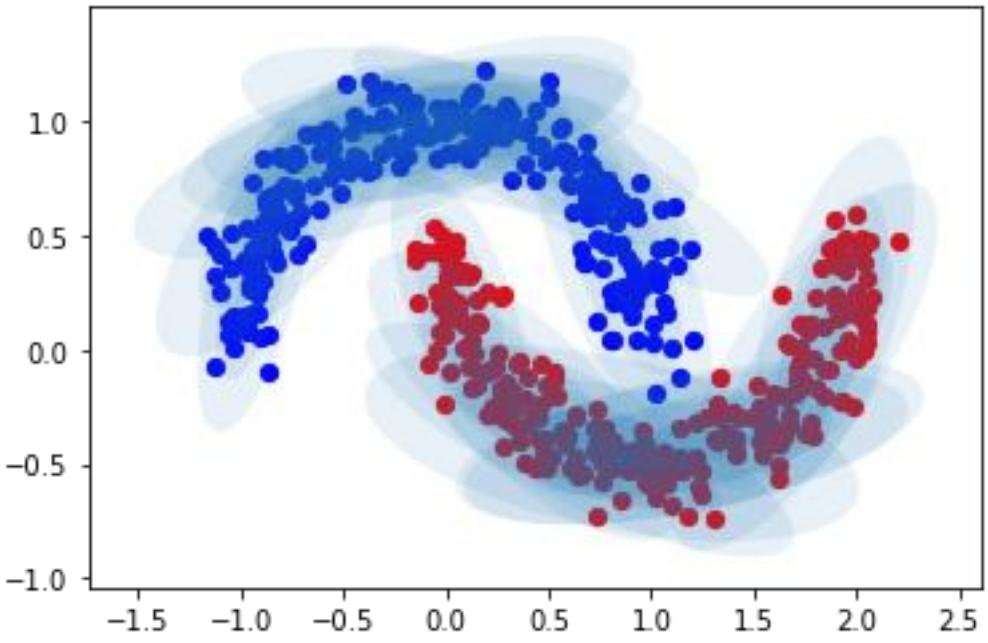


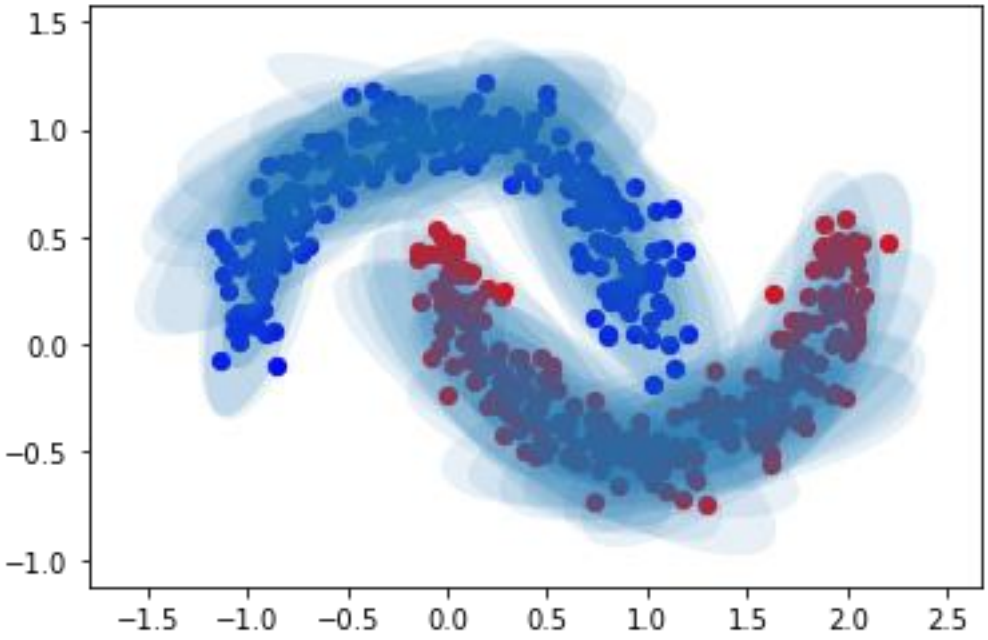


در کل میتوان تعداد پارامتر ۳ و یا ۶ را به عنوان بهترین تعداد کامپوننت انتخاب کرد.(بر اساس کمترین شدن مجموع aic و bic)

در نهایت هم نمودار برای تعداد کامپوننت ۳ و ۸و و۱۶ را می کشیم که آن ها را در زیر می بینیم:


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5


1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5