مراحل انجام پروژه

1. کتابخانه ها را بارگیری کرده و داده ها را بخوانید
   - 1.1 بارگیری کتابخانه ها
   - 1.2 داده ها را بخوانید
   - 1.3 مقادیر از دست رفته
   - 1.4 ویژگی های بی فایده را رها کنید

2. تجزیه و تحلیل اکتشافی داده ها (EDA)
   - 2.1 توصیف دیتاست
   - 2.2. توزیع هدف (diagnosis)
   - 2.3 توزیع ویژگی ها
   - 2.4 ماتریس همبستگی
   - 2.5 ویژگی های همبستگی مثبت
   - 2.6 ویژگی های غیر همبسته
   - 2.7 ویژگی های همبستگی منفی

3. تجزیه و تحلیل مولفه اصلیPCA
   - 3.1 PCAرا محاسبه کنید
   - 3.2 PCA با 6 جز
   - 3.3 نمودار پراکندگی PCA با 2 جز
   - 3.4 نمودار پراکندگی PCA با 3 جز

4. تعریف توابع

4.1 ماتریس سردرگمی

4.2 منحنی Precision-Recall

4.3 منحنیROC

4.4 منحنی یادگیری

4.5 متغیرهای اعتبار سنجی

5. مجموعه داده را آماده کنید

5.1 تعریف X, y

5.2 مقیاس کننده استاندارد(X)

5.3 تقسیم دیتاست به آموزش و تست

6. مدل پیش بینی اول: رگرسیون لجستیک Logistic Regression

6.1 و GridSearch CV برای بهینه سازی ابرپارامترها

6.2 حذف بازگشتی ویژگی های

6.3 مقایسه منحنی های یادگیری و نمرات اعتبار متقابل

مجموعه داده

breast cancer wisconsin

ویژگی ها از طریق یک تصویر دیجیتالی از یک آسپیرات سوزنی ظریف از توده پستان محاسبه می شوند. آنها ویژگی های هسته سلول موجود در تصویر را توصیف می کنند.

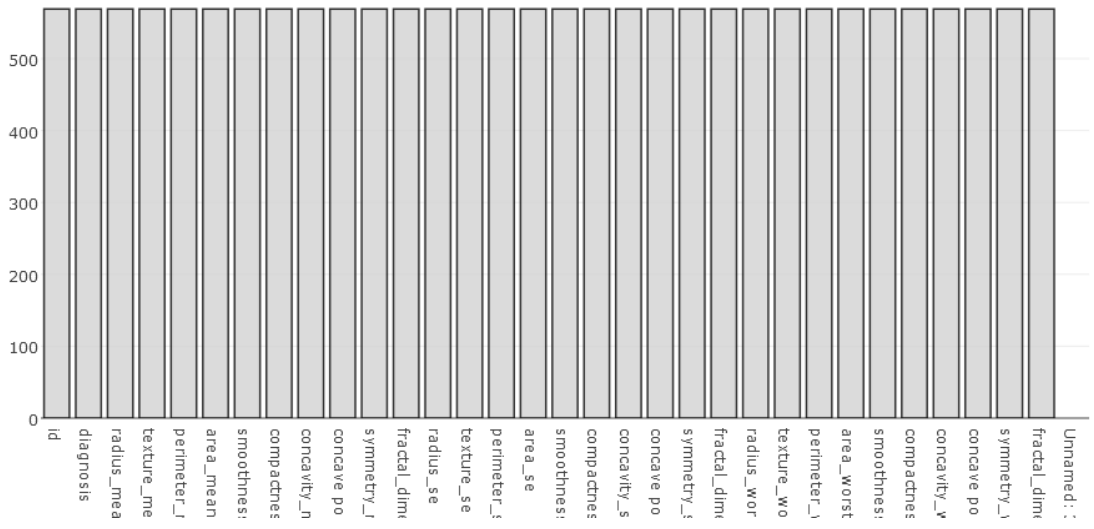
- شماره شناسایی ID
- تشخیص (Mبدخیم ، Bخوش خیم)

ده ویژگی با ارزش واقعی برای هر هسته سلول محاسبه می شود:

- شعاع (میانگین فاصله از مرکز تا نقاط محیط)
- بافت (انحراف معیار مقادیر خاکستری)
- محیط
- حوزه
- صافی (تغییر محلی در طول شعاع)
- فشردگی (محیط ^ 2 / مساحت - 1.0)
- تقعر (شدت قسمت های مقعر کانتور)
- نقاط مقعر (تعداد قسمت های مقعر کانتور)
- تقارن
- بعد فراکتال ("تقریب خط ساحلی" - 1)

مرحله اول

حذف مقادیر از دست رفته:



با توجه به نمودار تمام ویژگی ها کامل هستند به جز ستون Unnamed: 32 که به کلی null است که آن را حذف می کنیم

مرحله دوم

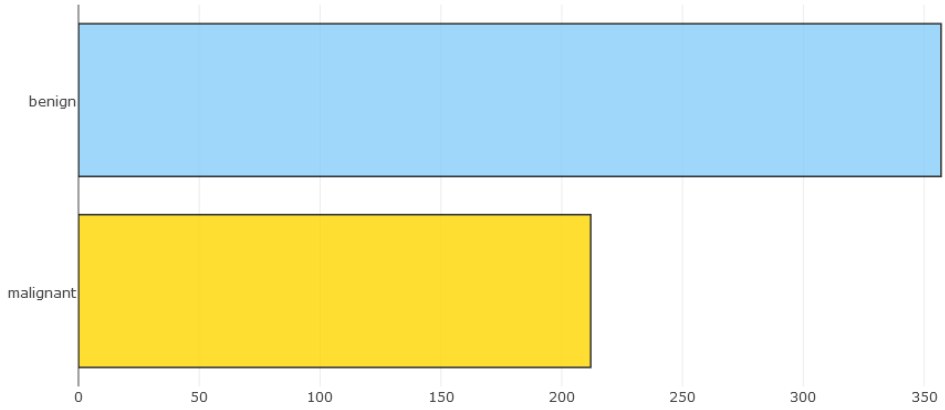
تجزیه و تحلیل اکتشافی داده ها (EDA)

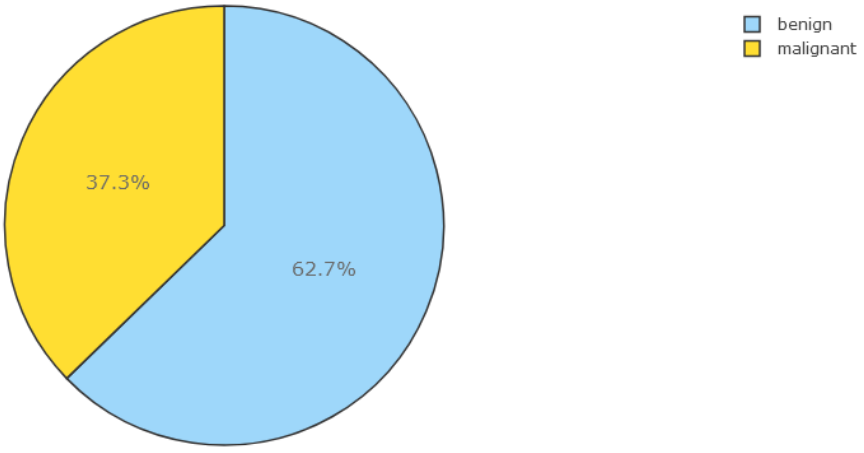
تحلیل اکتشافی داده ها یا EDA ، مرحله اول و مهم در تحلیل هر گونه داده است. اهداف اصلی تحلیل اکتشافی عبارتند از:

- شناسایی اشتباهات

- بررسی فرضیات
- انتخاب اولیه مدل‌های مناسب
- تعیین روابط بین متغیرهای کمکی (تبیینی، توضیحی، مستقل)
- ارزیابی اولیه میزان و جهت ارتباط میان متغیرهای مستقل و وابسته

توزیع هدف (تشخیص)

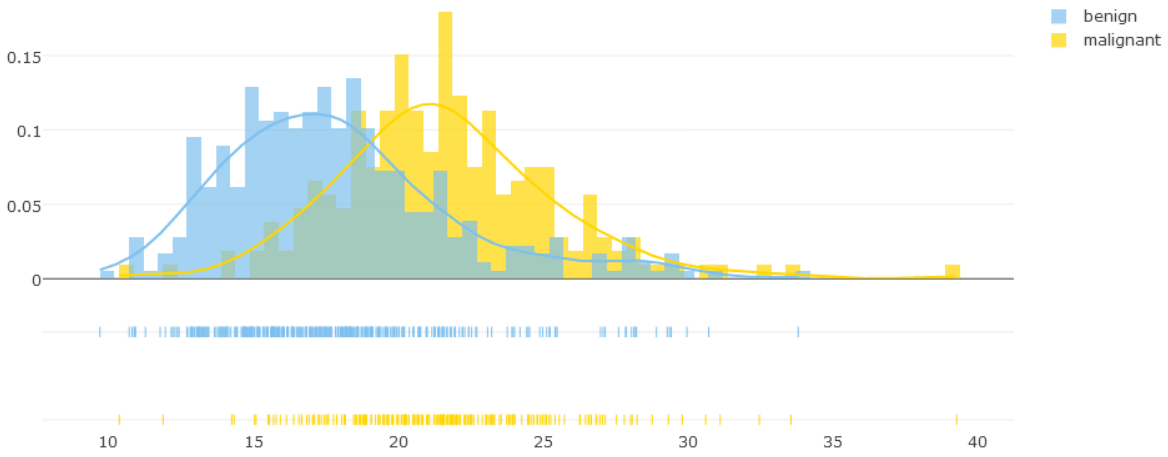




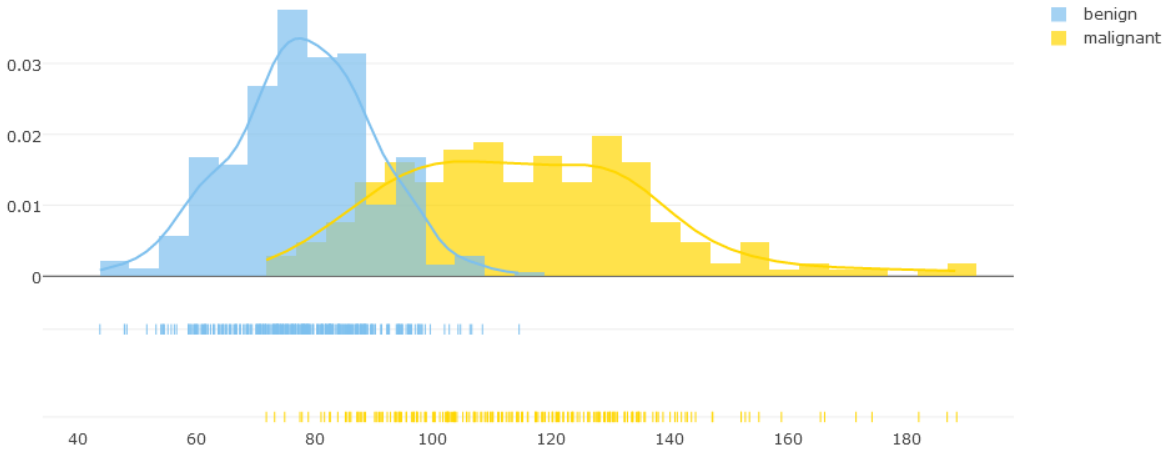
با توجه به نمودار ها در میابیم که مقادیر خوش خیم درصد بیشتری از تشخیص ها را به خود اختصاص داده اند

توزیع ویژگی ها(میانگین)

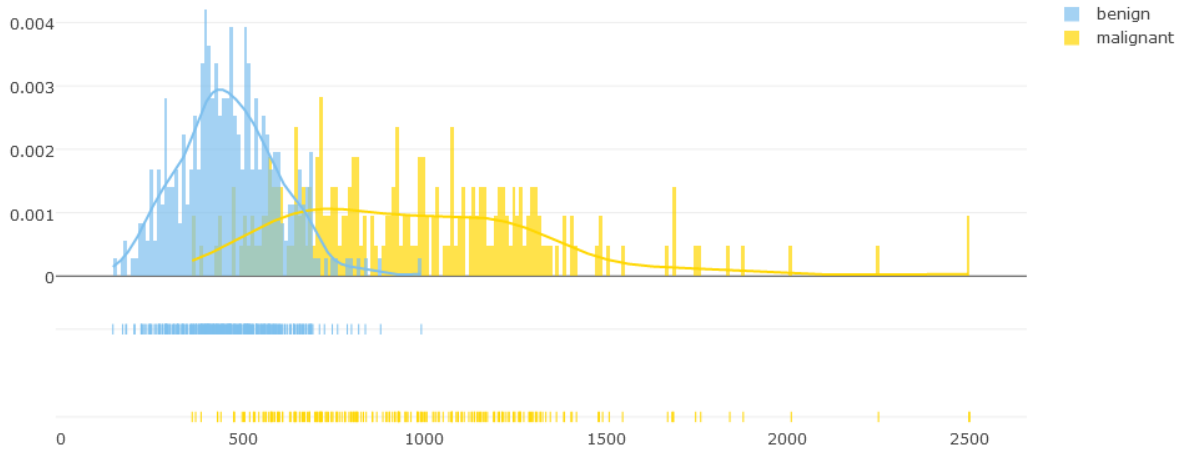
Texture



Perimeter



Area



ماتریس همبستگی

ماتریس همبستگی (Correlation Matrix)جدولی است که ضریب همبستگی بین متغییر ها را نشان می دهد. هر متغیر تصادفی مثل X_i در جدول با هریک از متغییرهای دیگر (X_j)همبستگی دارد. با این جدول می توانیم بفهیم کدام متغیرها بیشترین همبستگی را با هم دیگر دارند.

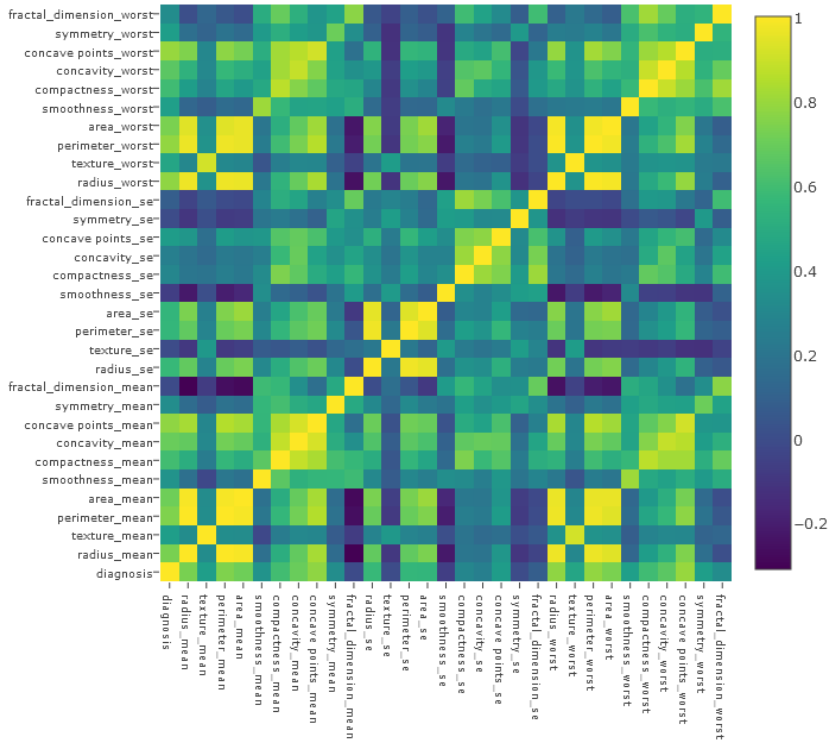
وقتی می گوییم دو متغیر همبستگی دارند یعنی تغییر یکی از متغییرها بر مقدار پارامتر دیگری تاثیر می گذارد. یا به عبارت دیگر تغییر دو پارامتر وابسته به هم است.

- اگر ضریب همبستگی دو پارامتر با یکدیگر مثبت باشد، به این معناست که در فضایی که مطالعه و بررسی انجام شده، افزایش یک پارامتر با افزایش پارامتر دیگر و نیز کاهش آن پارامتر با کاهش پارامتر دیگر همراه است .

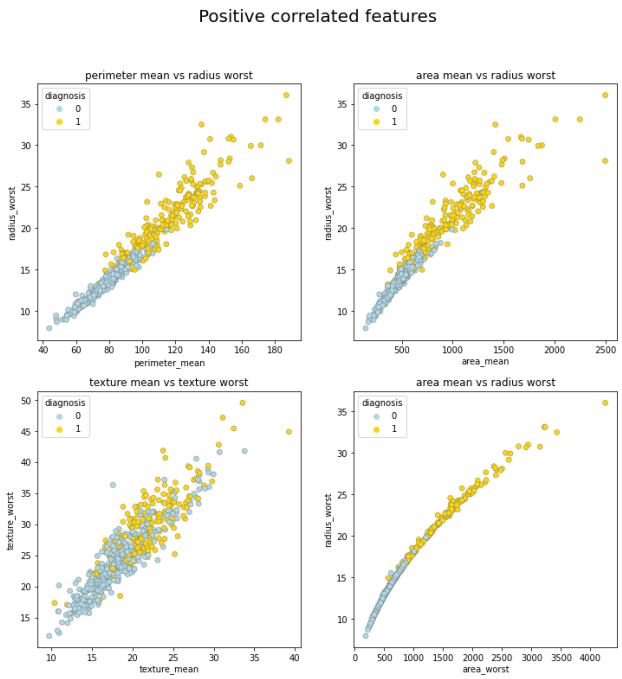
- اگر ضریب همبستگی دو پارامتر با یکدیگر منفی باشد، به این معناست که در فضایی که مطالعه و بررسی انجام شده، افزایش یک پارامتر با کاهش پارامتر دیگر و کاهش آن پارامتر با افزایش پارامتر دیگر همراه است .

- صفر بودن ضریب همبستگی به این معناست که دو پارامتر – در فضایی که مورد بررسی قرار گرفته – مستقل از یکدیگر بوده‌اند و بر اساس اطلاعات موجود از کاهش یا افزایش یکی، نمی‌توان در مورد کاهش یا افزایش دیگری اظهار نظر کرد .

- ضریب همبستگی بین منفی یک و مثبت یک است. هر چه این ضریب از صفر دورتر شود (و به مثبت یا منفی یک نزدیک‌تر شود) می‌توان نتیجه گرفت که روند هم جهت بودن یا مخالف بودن دو پارامتر مورد بررسی، جدی‌تر است.

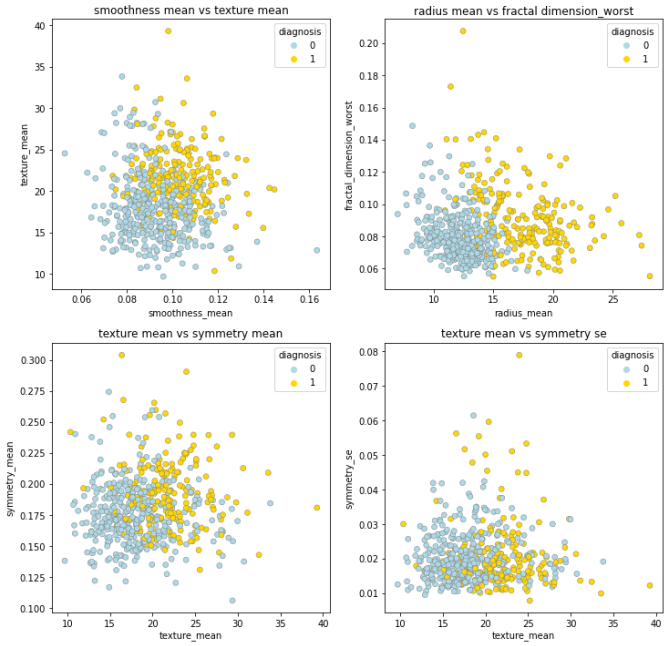


ویژگی های همبستگی مثبت



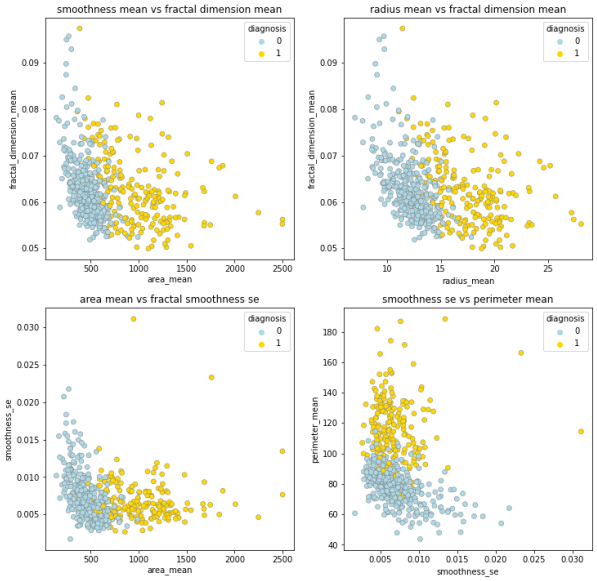
ویژگی های غیر همبسته

Uncorrelated features



ویژگی های همبستگی منفی

Negative correlated features

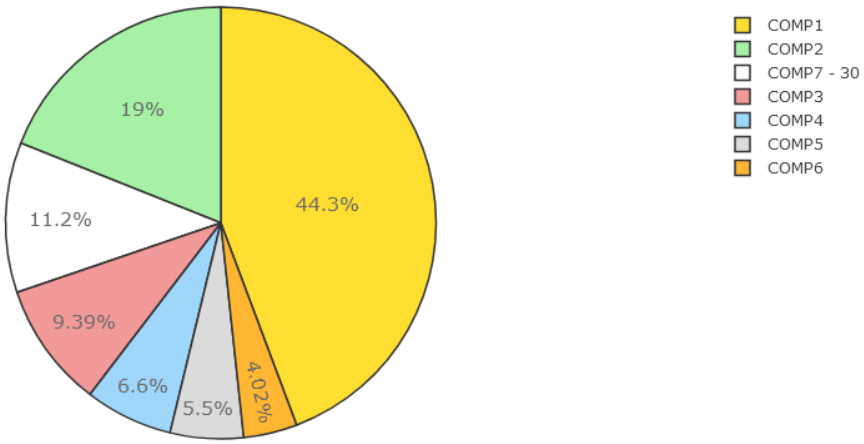


مرحله سوم

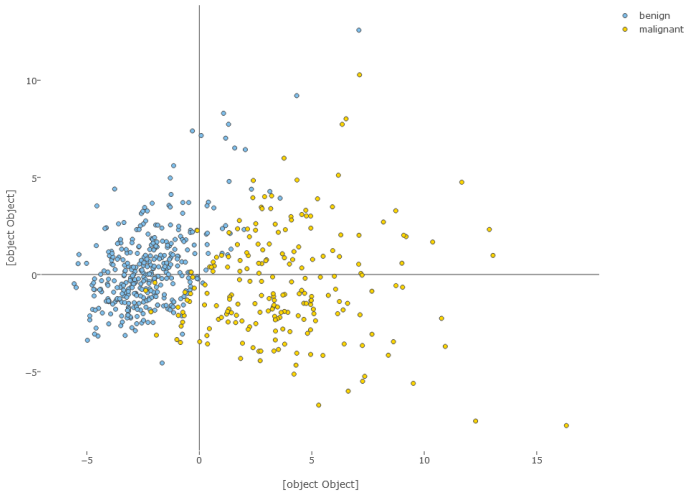
تجزیه و تحلیل مولفه اصلیPCA :

تحلیل مولفه‌های اصلی (Principal Component Analysis - PCA) تبدیلی در فضای برداری است، که بیشتر برای کاهش ابعاد مجموعهٔ داده‌ها مورد استفاده قرار می‌گیرد. تحلیل مولفه‌های اصلی در تعریف ریاضی یک تبدیل خطی متعامد است که داده را به دستگاه مختصات جدید می‌برد به طوری که بزرگترین واریانس داده بر روی اولین محور مختصات، دومین بزرگترین واریانس بر روی دومین محور مختصات قرار می‌گیرد و همین طور برای بقیه. تحلیل مولفه‌های اصلی می‌تواند برای کاهش ابعاد داده مورد استفاده قرار بگیرد، به این ترتیب مولفه‌هایی از مجموعه داده را که بیشترین تاثیر در واریانس را دارند حفظ می‌کند.

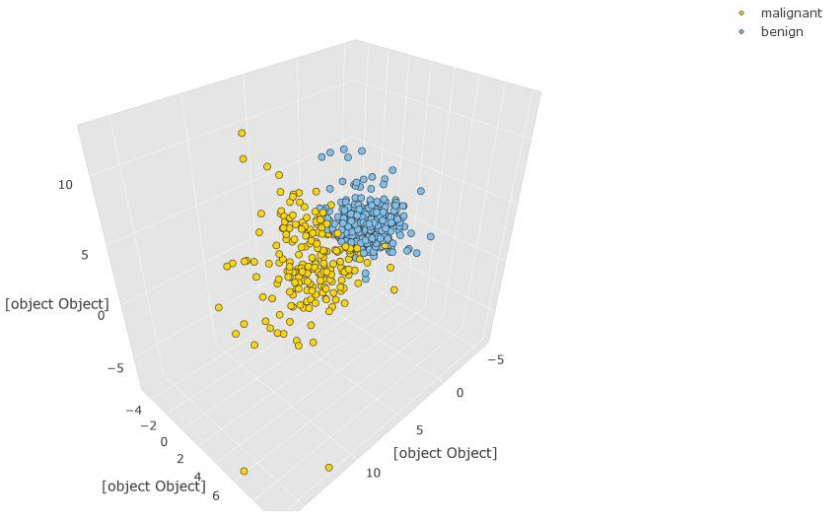
PCAبا 6 جز



نمودار پراکندگی PCA با 2 جز



نمودار پراکندگی PCA با 3 جز



مرحله چهارم

تعریف توابع

ماتریس سردرگمی ، همچنین به عنوان ماتریس خطا شناخته می شود ، امکان تجسم عملکرد یک الگوریتم را فراهم می کند:

- مثبت مثبت (TP) : تومور بدخیم به درستی بدخیم شناخته شده است
- منفی واقعی (TN) : تومور خوش خیم به درستی خوش خیم شناخته شده است
- مثبت کاذب (FP) : تومور خوش خیم به اشتباه به عنوان بدخیم شناخته شده است
- منفی کاذب (FN) : تومور بدخیم به اشتباه به عنوان خوش خیم شناخته شده است

Accuracy : (TP +TN) / (TP + TN + FP +FN)

Precision : TP / (TP + FP)

Recall : TP / (TP + FN)

- منحنی Precision-Recall معامله بین دقت و فراخوان برای آستانه متفاوت را نشان می دهد
- منحنی ROC با ترسیم نرخ مثبت واقعی TPR در برابر نرخ مثبت کاذب FPR در تنظیمات مختلف آستانه ایجاد می شود.
- منحنی یادگیری ، نمرات آموزش و آزمون با اعتبار متقابل را تعیین می کند.
- اعتبارسنجی متقابل روشی برای ارزیابی مدلهای پیش بینی از طریق تقسیم نمونه اصلی در یک مجموعه آموزشی برای آموزش مدل و یک مجموعه آزمون برای ارزیابی آن است. منحنی یادگیری ، نمرات آموزش و آزمون با اعتبار متقابل تعیین می کند.

مرحله پنجم

آماده کردن مجموعه داده

y = (هدف) تشخیص

X = (شعاع_معنا ، ناحیه_س ، ....) ویژگی ها

در گام بعدی، داده‌ها به دو دسته آموزش (Train) و تست (Test) تقسیم می‌شوند که در آن‌ها، مجموعه تست ۲۰ درصد از داده‌ها را به خود اختصاص می‌دهد. در اینجا، از random_state برای حصول اطمینان از این موضوع استفاده می‌شود که تقسیم‌بندی در هر بار ثابت می‌ماند.

```
#Train_test split
random_state = 42
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.20, random_state = random_state)
```

استاندارد سازی داده ها

چرا لازم است داده های خود را استاندارد کنید؟ به عنوان مثال ، متغیری که بین 0 تا 100 باشد ، از متغیری که بین 0 تا 1 باشد ، وزن بیشتر خواهد داشت. استفاده از این متغیرها بدون استاندارد بودن ، وزن بیشتری در تجزیه و تحلیل به متغیر با دامنه بزرگتر می دهد.

مرحله ششم

مدل پیش بینی اول: رگرسیون لجستیک Logistic Regression و GridSearch CV برای بهینه سازی ابرپارامترها

رگرسیون لجستیک

رگرسیون لجستیک دووجهی Binomial Logistic Regression تکنیک آماری رگرسیونی است که به ما امکان مدل بندی داده های کیفی دودویی) Dichotomous وقوع یا عدم وقوع پیشامد) به عنوان متغیر پاسخ را میدهد و امروزه در تمام زمینه های علمی کاربرد گسترده ای یافته است. در مباحث مطرح شده در رگرسیون خطی متغیرهای پیشگو می توانند هم کمی و هم کیفی باشند ولی عموما متغیر پاسخY، متغیری کمی و پیوسته در نظر گرفته می شود.

بهینه سازی ابرپارامترها

همانطور که می دانیم مدل های Ml به گونه ای دارای پارامتر هستند که رفتار آنها برای یک مساله خاص می تواند تعدیل شود. تنظیم الگوریتم به معنی پیدا کردن بهترین ترکیب از این پارامترها است به گونه ای که کارایی مدل ML بهبود یابد. این روند گاهی بهینه سازی فرا پارامتر (hyperparameter)نام می گیرد ،و پارامتر های خود الگوریتم فرا پارامتر (hyperparameter)نام دارند و ظرایب الگوریتم ML پارامتر نام می گیرد.

در اینجا، به بحث درباره برخی از متدها برای تنظیم پارامتر الگوریتم که توسط Python Scikit-learn ارائه شده است، خواهیم پرداخت.

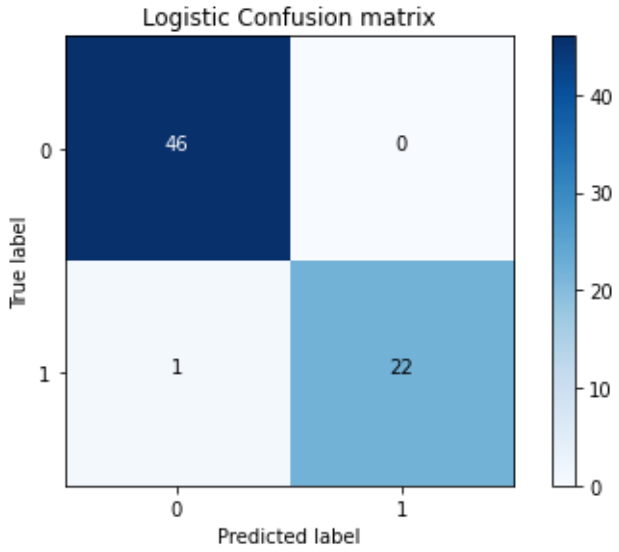
تنظیم پارامتر جستجوی شبکه توری (grid search)

این یک راهکار تنظیم پارامتر است. نقطه کلیدی درباره عملکرد این متد این است که، برای هر ترکیب ممکن از پارامترهای الگوریتم، که در شبکه توری مشخص شده باشد، مدل را به صورت روشمند ساخته و ارزیابی می کند. از این رو ، می توان گفت که این الگوریتم ماهیت جستجو دارد.

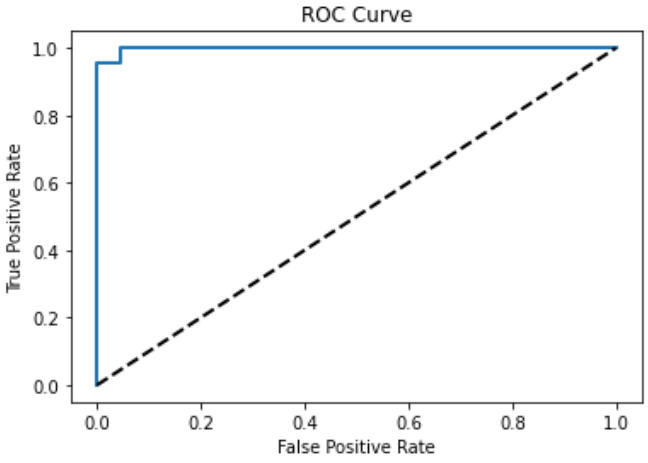
نتایج:

Accuracy = 0.986
Precision = 1.000
Recall = 0.957
F1_score = 0.978

ماتریس آشفتگی



نمودار ROC

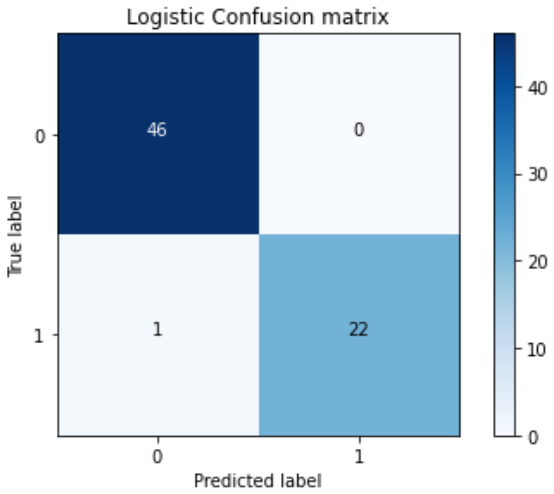


حذف بازگشتی ویژگی های

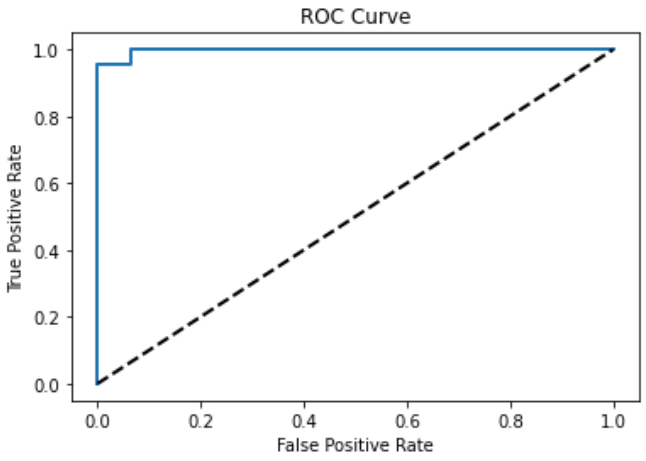
روش «حذف بازگشتی ویژگی(Recursive Feature Elimination) «، یک روش »حریصانه (Greedy) «برای انتخاب ویژگی است. در این روش، ویژگی‌ها به طور بازگشتی و با در نظر گرفتن مجموعه‌های کوچک و کوچک‌تر از ویژگی‌ها (در هر مرحله) انتخاب می‌شوند. در این روش، ویژگی‌ها بر اساس مرتبه حذف شدن آن‌ها از فضای ویژگی رتبه‌بندی می‌شوند. در زبان برنامه برنامه‌نویسی پایتون، از بسته نرم‌افزاری SciKit-Learnبرای حذف بازگشتی ویژگی‌های نامرتبط و انتخاب بهترین ویژگی‌ها استفاده می‌شود .

نتایج

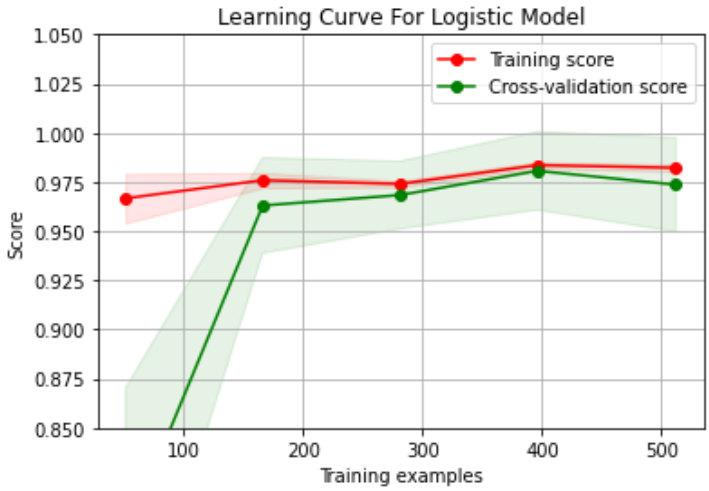
ماتریس آشفتگی



نمودار ROC

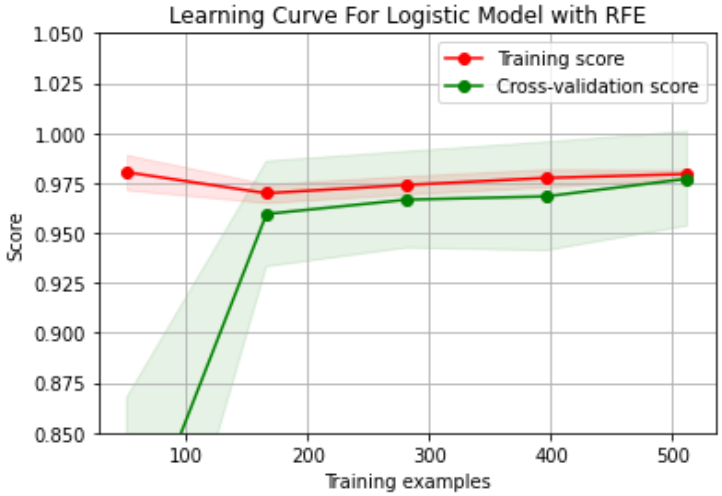


مقایسه منحنی های یادگیری و نمرات اعتبار متقابل



1-

[accuracy] : 0.97541 (+/- 0.00653)
[precision] : 0.99024 (+/- 0.01196)
[recall] : 0.94352 (+/- 0.01856)



2-

[accuracy] : 0.97365 (+/- 0.01107)
[precision] : 0.98524 (+/- 0.01206)
[recall] : 0.94352 (+/- 0.02368)