Homework 6

import numpy as np

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.decomposition import PCA

from sklearn.preprocessing import scale

data=pd.read\_csv("D:/Old\_Data/math/Data science toseeh/Files/Spine.csv")

data.shape

output: (310,7)

این داده دارای 310 نمونه با 7 ویژگی زیر است.

میزان لغزش مهره کمر یا GS

lumbar lordosis angle زاویه لوردوز کمر

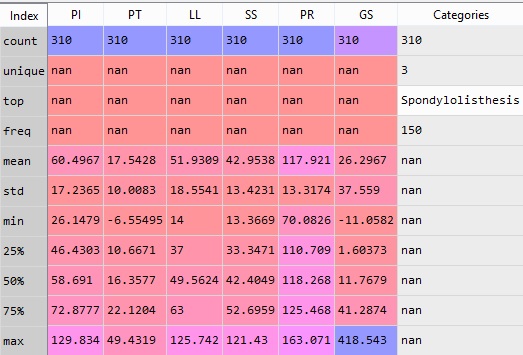
pelvic tilt شیب لگن

sacral slope شیب ساکارال

pelvic radius شعاع لگن

pelvic incidence

data.describe(include="all")



با توجه به اینکه مقدار count برای همه برابر با 310 است پس داده گمشده نداریم.

data.dtypes

تمام مغیر ها یه غیر از متغیر پاسخ کمی هستند. متغیر پاسخ کیفی است با 3 کلاس نرمال، انحراف ستون فقرات و دیسک کمر.

راه دیگر برای مشاهده مقادیر گمشده در هر یک از ویژگی ها.

np.sum(data.isna())

PI 0

PT 0

LL 0

SS 0

PR 0

GS 0

Categories 0

dtype: int64

c)

data.agg({"PI":("mean","std"),"GS":("mean","median","std")})

PI GS

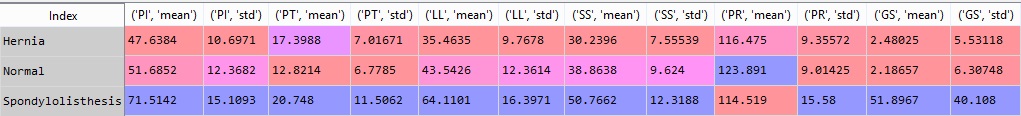
mean 60.496653 26.296694

median NaN 11.767934

std 17.236520 37.559027

d)

a=data.groupby("Categories").agg(("mean","std"))



محاسبه فاصله 3 برابر انحراف معیار و بازه اطمینان برای هر یک از ویژگی ها در هر کلاس.

len(data[data["Categories"]=="Hernia"])

out:60

len(data[data["Categories"]=="Normal"])

out:100

len(data[data["Categories"]=="Spondylolisthesis"])

out:150

#PI-Hernia

فرمول ها در ورد جا به جا شده. این فرمول محاسبه بازه اطمینان است.

47.6384+1.96\*(10.6971/np.sqrt(60))

Out: 50.345139756644066

47.6384-1.96\*(10.6971/np.sqrt(60))

Out: 44.93166024335593

بنابراین بازه اطمینان این ویژگی در کلاس دیسک کمر برابر با (44.931,50.345) است

#PI-Normal

51.6852+1.96\*(12.3682/np.sqrt(100))

Output: 54.1093672

51.6852-1.96\*(12.3682/np.sqrt(100))

Output: 49.2610328

بنابراین بازه اطمینان این ویژگی در کلاس نرمال برابر با (49.261,54.109) است

#PI-Spondy

71.5142+1.96\*(15.1093/np.sqrt(100))

Output: 73.93219159088142

71.5142-1.96\*(15.1093/np.sqrt(100))

Output: 69.09620840911859

بنابراین بازه اطمینان این ویژگی در کلاس لغزی ستون فقرات برابر با (69.09,73.93) است.

پس اگرنمونه گیری را تعداد دفعات زیادی انجام دهیم 95 درصد مواقع میانگین این ویژگی در این کلاس در فاصله بیان شده است و چون با دو بازه دیگر اشتراک ندارد پس به طور اماری میانگین این کلاس بالاست.

حال نمودار هیستوگرام این ویژگی را برای کلاس های متفاوت رسم میکنیم.

a1=np.mean(data["PI"].loc[data["Categories"]=="Normal"])

a2=np.mean(data["PI"].loc[data["Categories"]=="Hernia"].dropna())

a3=np.mean(data["PI"].loc[data["Categories"]=="Spondylolisthesis"])

grid=sns.FacetGrid(hue="Categories",data=data)

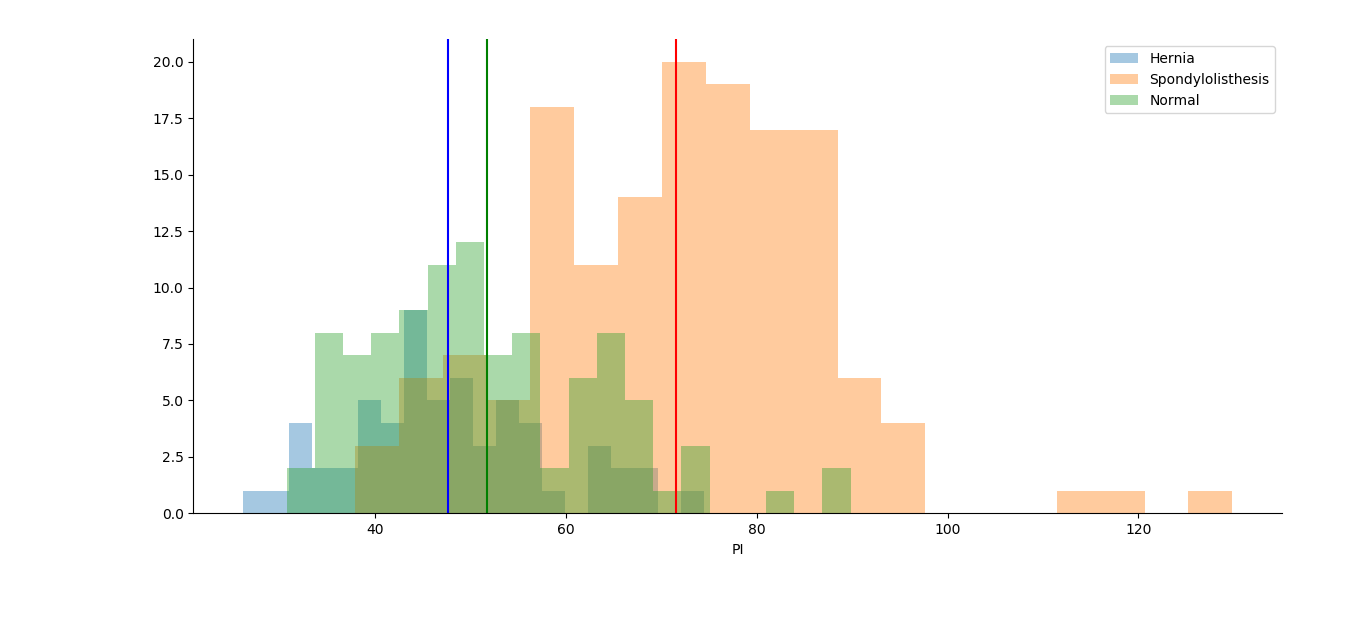
grid.map(plt.hist,"PI",alpha=0.4,bins=20)

plt.legend()

plt.axvline(a1,color='green')

plt.axvline(a2,color='blue')

plt.axvline(a3,color='red')



با توجه به نمودار با داشتن تنها این ویژگی دو کلاس نرمال و دیسک از هم قابل تشخیص نیستند اما تا حدودی میتوان کلاس لغزش ستون فقرات را مشخص کرد. اگر داده اموزشی دارای مقدار ویژگی PI در بازه حدودا 95 به بعد باشد صد در صد دارای لغزش ستون فقرات است. اما کمتر از این مقدار را نمیتوان تشخص داد و ویژگی های دیگری نیاز دارد.

دو داده ای که دارای مقدار حدودا 118 و 129 در کلاس لغزش هستند داده پرت در این ویژگی هستند چون از دیگر نقاط در این ویژگی که در کلاس لغزش هستند فاصله دارند و تعداد انها زیاد نیست که بتوان پرت حساب نکردو به عنوان یک گروه در نطر گرفت. (مقدار انها از 71.51+3\*15.10 بیشتر است.)

data[data["PI"]>=116]

Out[63]:

PI PT ... GS Categories

115 129.834041 8.404475 ... 418.543082 Spondylolisthesis

162 118.144655 38.449501 ... 74.043767 Spondylolisthesis

ویژگی PT:

بازه های اطمینان به صورت زیر هستند. توجه در فرمول قبلی میانگین و انحراف معیار این ویژگی در کلاس های متفاوت جایگزین شده است.

بازه اطمینان کلاس دیسک کمر. (15.62,19.17)

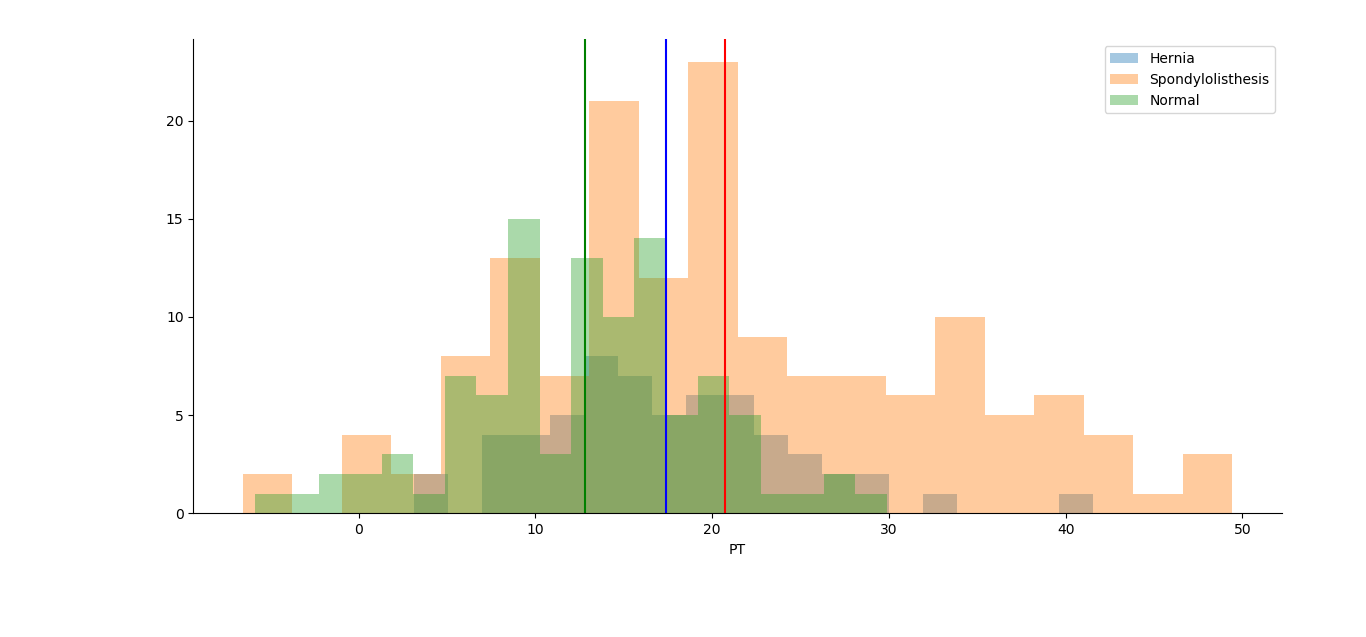
بازه اطمینا این ویژگی در کلاس نرمال (11.49,14.14)

بازه اطمینان در کلاس لغزش ستون فقرات (18.90,22.58)

در این ویژگی هم اگر نمونه گیری را زیاد انجام دهیم در 95 درصد مواقع میانگین این ویژگی در کلاس لغزش در بازه اطمینان داده شده است و چون با دیگر بازه ها اشتراک ندارد این میانگین در این کلاس از نظر اماری زیاد است.

برای دو بازه دیگر هم به همین صورت است. هر سه بازه دو به دو اشتراک تهی دارند. پس میانگین ها از نظر اماری متفاوت هستند. اما برای اینکه متوجه شویم که ایا این ویژگی به تنهایی میتوناد متغیر پاسخ را توصیف کند هیستوگرام را رسم میکنیم.

(دستورها مشابه قبل هیتند اما برای ویژگی PT .)



بنابر این هیستوگرام این ویژگی به تنهایی کلاس ها را نمیتواند از هم متمایز کند.

بدون نمودار میتوان mean-3\*std,mean+3\*std) ) را محاسبه نمود.

ویژگی LL:

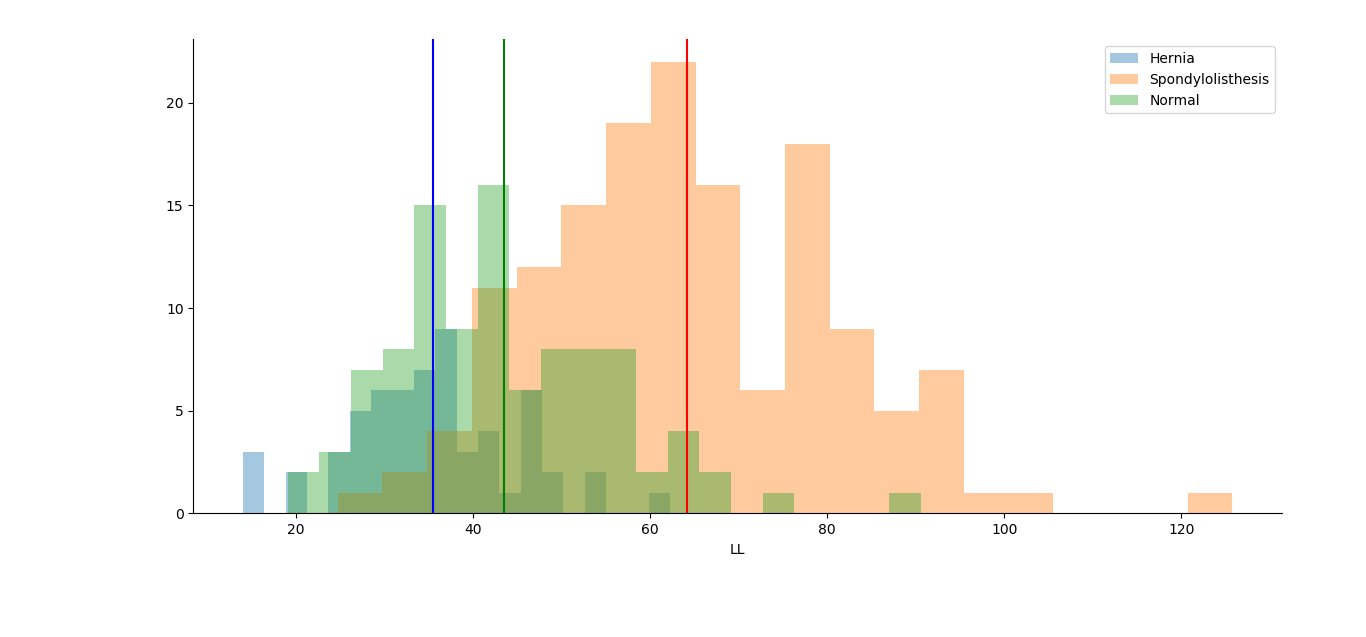
بازه های اطمینان به صورت زیر هستند. توجه در فرمول قبلی میانگین و انحراف معیار این ویژگی در کلاس های متفاوت جایگزین شده است.

بازه اطمینان کلاس دیسک کمر. (32.99,37.92)

بازه اطمینا این ویژگی در کلاس نرمال (41.11,45.96)

بازه اطمینان در کلاس لغزش ستون فقرات (61.48,66.73)

هر سه بازه دو به دو اشتراک تهی دارند. پس میانگین ها از نظر اماری متفاوت هستند. اما برای اینکه متوجه شویم که ایا این ویژگی به تنهایی میتواند متغیر پاسخ را توصیف کند هیستوگرام را رسم میکنیم.

داده ای که دارای مقدار حدودا 125 است در این ویژگی داده ای پرت است زیرا از (64.11+3\*16.39) بیشتر است.

data[data["LL"]>=113]

Out[65]:

PI PT ... GS Categories

197 58.828379 37.577873 ... 117.314683 Spondylolisthesis

ویژگی SS:

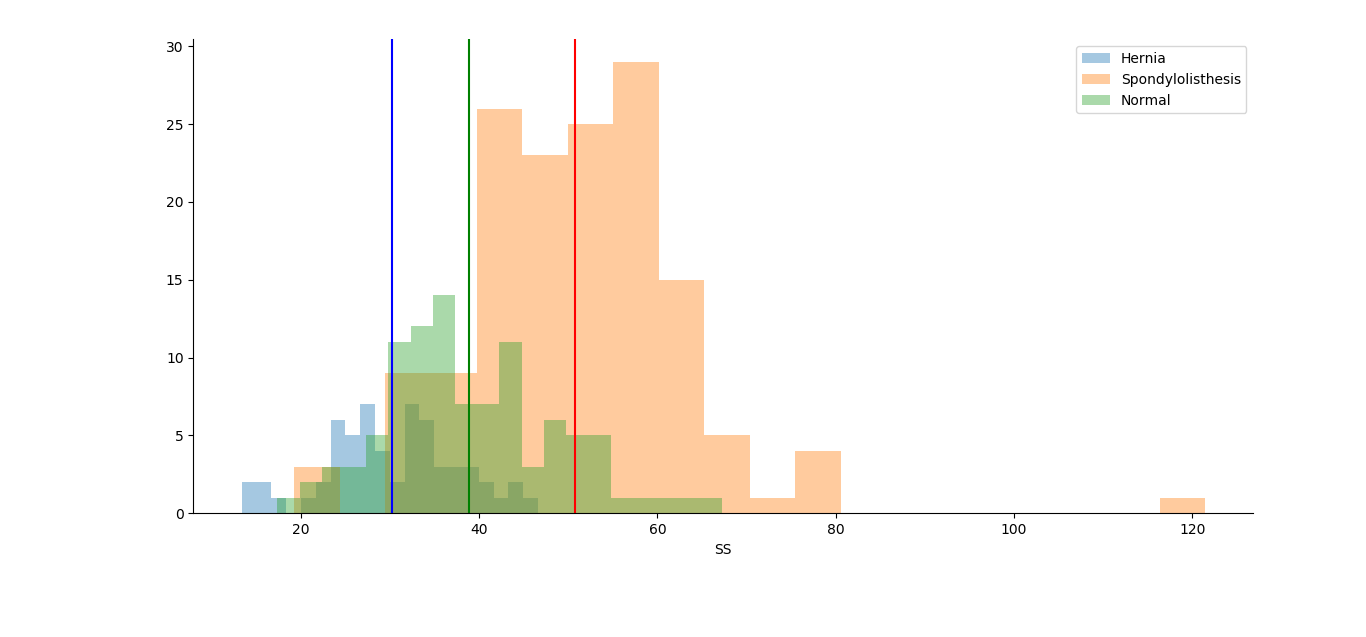
بازه های اطمینان به صورت زیر هستند. توجه در فرمول قبلی میانگین و انحراف معیار این ویژگی در کلاس های متفاوت جایگزین شده است.

بازه اطمینان کلاس دیسک کمر. (28.31,32.14)

بازه اطمینا این ویژگی در کلاس نرمال (36.97,40.74)

بازه اطمینان در کلاس لغزش ستون فقرات (48.78,52.73)

هر سه بازه دو به دو اشتراک تهی دارند. پس میانگین ها از نظر اماری متفاوت هستند.



دوکلاس نرمال و دیسک را نمیتواند متمایز کند ولی بسته به مقدار این ویژگی شاید بتوان کلاس لغزش را مشخص نمود. میانگین ها از نطر اماری متفاوت هستند اما چون انحراف معیار دو کلاس نرمال و دیسک زیاد استد این سه کلاس خیلی از هم قبال تشخیص نیستند.

ویژگی PR:

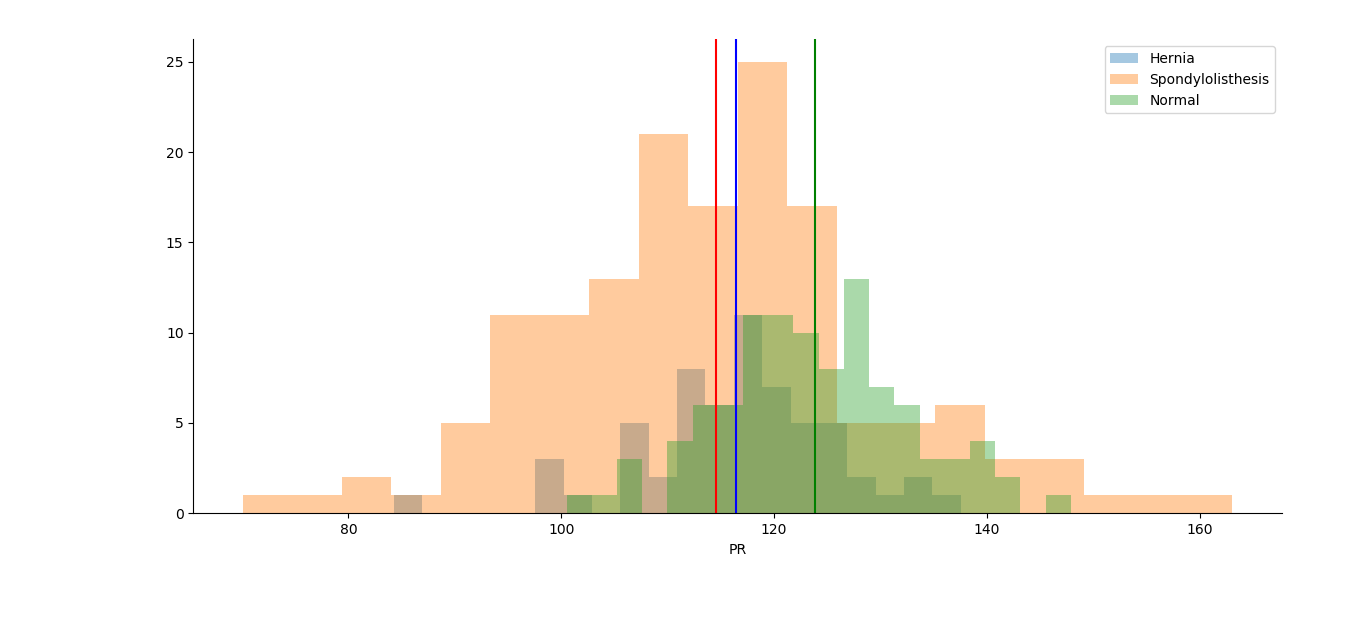
بازه های اطمینان به صورت زیر هستند. توجه در فرمول قبلی میانگین و انحراف معیار این ویژگی در کلاس های متفاوت جایگزین شده است.

بازه اطمینان کلاس دیسک کمر. (114.10,118.83)

بازه اطمینا این ویژگی در کلاس نرمال (122.12,125.65)

بازه اطمینان در کلاس لغزش ستون فقرات (112.01,117)

میانگین کلاس نرمال از لحاظ اماری از دو کلاس دیگر بیشتر است.



به نظر من سه کلاس از یکدیگر قابل تشخیص نیستند و داده پرتی هم نداریم در کل. اما برای بررسی که ایا در کلاس نرمال و دیسک داده پرت داریم یا نه:

data[(data["PR"]>=151)&(data["Categories"]=="Normal")]

Out[90]:

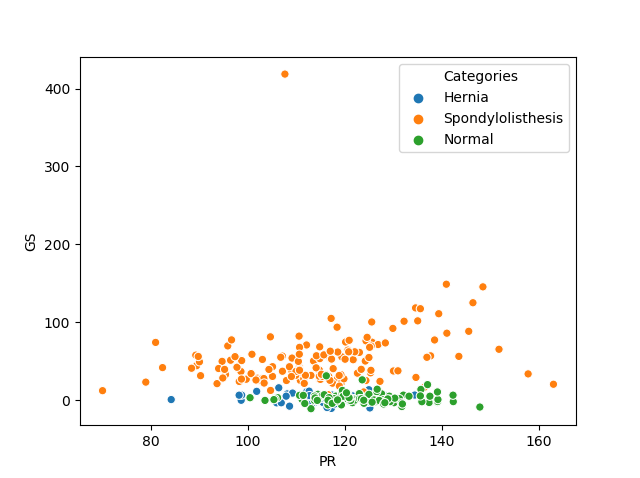
Empty DataFrame

Columns: [PI, PT, LL, SS, PR, GS, Categories]

np.mean(data["PI"].loc[data["Categories"]=="Hernia"])-3\*np.std(data["PI"].loc[data["Categories"]=="Hernia"])=15.81

np.mean(data["PI"].loc[data["Categories"]=="Hernia"])+3\*np.std(data["PI"].loc[data["Categories"]=="Hernia"])

Out[57]: 79.46124827620135



داده ای که مقدار این ویژگی برای آن حدودا 90 است در کلاس دیسک کمر به نظر میرسد پرت باشد در این کلاس اما در بازه سه برابر انحارف معیار قرار دارد.

data[(data["PR"]<=88.42)&(data["Categories"]=="Hernia")]

PI 36.6864

PT 5.01088

LL 41.9488

SS 31.6755

PR 84.2414

GS 0.664437

Categories Hernia

ویژگی GS:

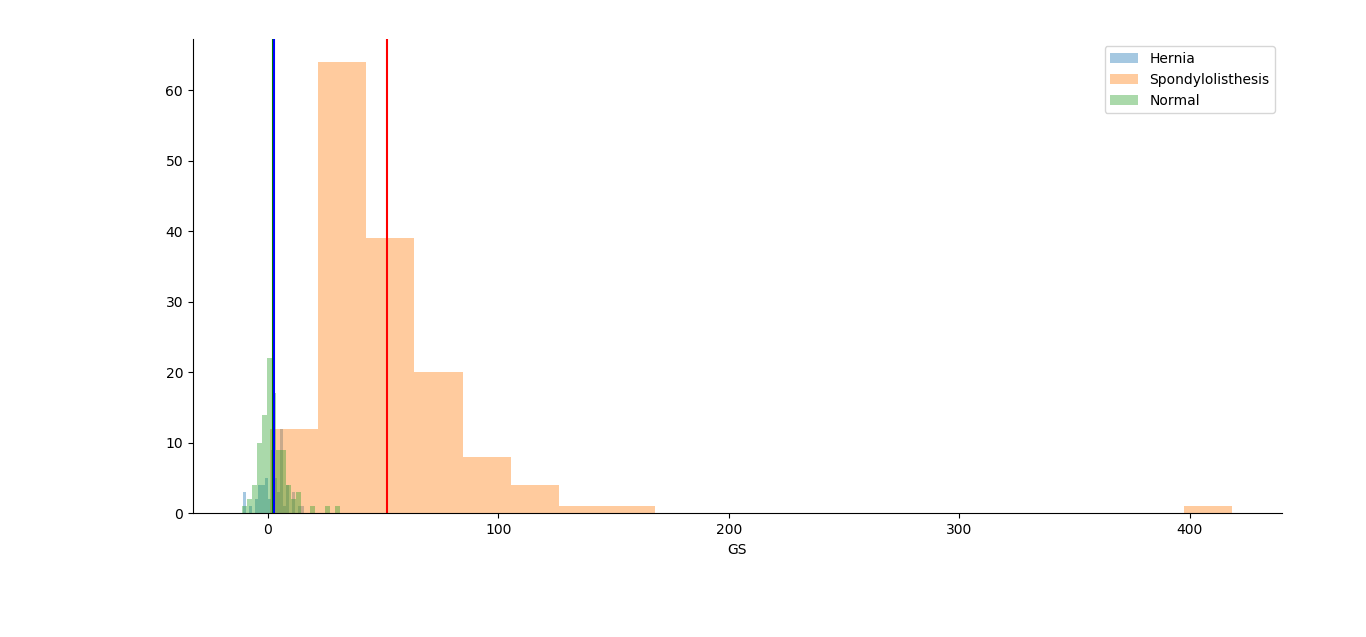
بازه های اطمینان به صورت زیر هستند. توجه در فرمول قبلی میانگین و انحراف معیار این ویژگی در کلاس های متفاوت جایگزین شده است.

بازه اطمینان کلاس دیسک کمر. (1.08,3.87)

بازه اطمینا این ویژگی در کلاس نرمال (0.94,3.41)

بازه اطمینان در کلاس لغزش ستون فقرات (45.47,58.30)

پس میانگین کلاس لغزش از نظر اماری از میانگین دو کلاس دیگر بیشتر است.

این ویژگی به خوبی کلاس لغزش ستون فقرات را تشخیص میدهد پس در داده های اموزشی اگر مقار این ویژگی از 40 بیشتر باشد حتما مربوط به کلاس لغزش است. اگر داده های جدیدی دهند و میانگین انها در بازه اطمینان گفته شده باشد (چون اگر این روند را ادامه دهیم 95 درصد مواقع میانگین در بازه اطمینان گفته شده است) در نتیجه داده ای هست که در کلاس لغزش است.

حتما داده ای وجود دارد که از میانگین بیشتر باشد و در نتیجه ان داده را به راحتی میتوان گفته در کلاس لغزش است.

با توجه به نمودار داده با مقدار GS برابر با 400 داده پرت است زیرا از سه برابر انحراف معیار یعنی یرا از

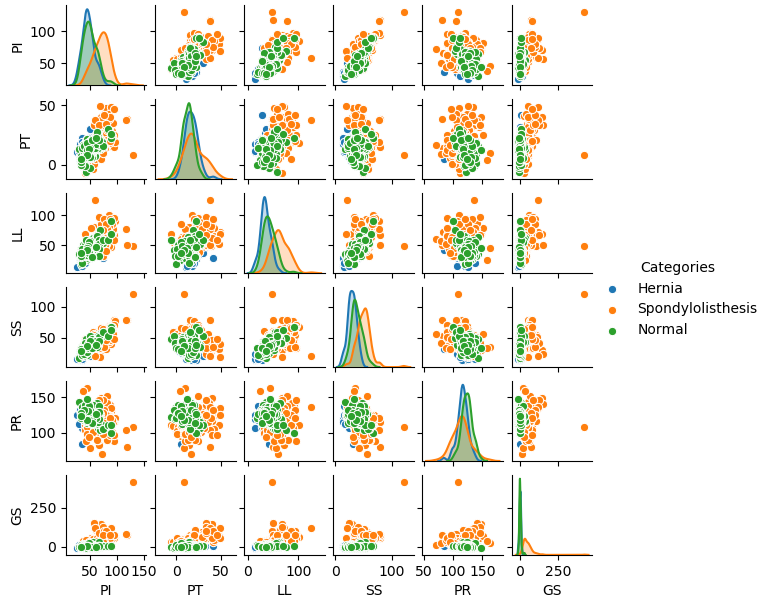
(51.89+3\*40.10) بیشتر است.

data[data["GS"]>=173]

Out[69]:

PI PT ... GS Categories

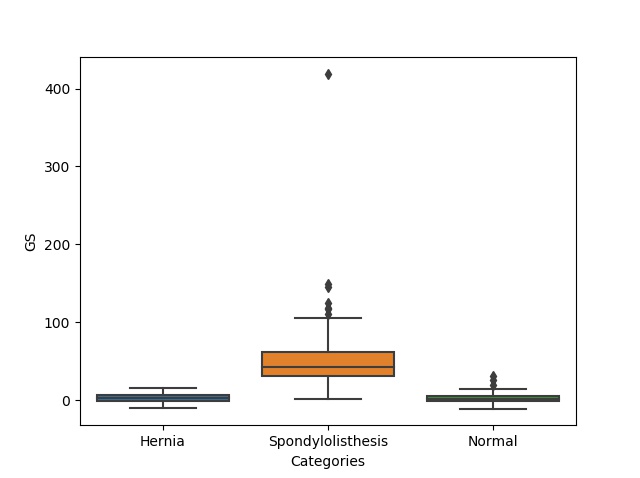
115 129.834041 8.404475 ... 418.543082 Spondylolisthesis



نقاط مشخص شده در شکل با مربع نقاطی هستند که در ویژگی PI پرت هستند زیرا این ویژگی دارای میانگین حدودا 60 و انحراف معیار حدودا 17.24 است اما این سه نقطه در کلاس لغزش ستون فقرات هستند و فقط دو تا از این نقاط یعنی داده 115 و 162 نقاط پرت در این کلاس هستند.

e)

sns.boxplot(x="Categories", y="GS", data=data)



میبینیم که میزان میانه در این ویژگی برای گروهی که دارای احنراف ستون فقرات هستند بیشتر است همچنین دارای انحراف معیار بیشتری هستند. اما این ویژگی برای متمایز کردن دیسک و نرمال خیلی مناسب نیست چون دارای میانه و و ماکزیمم و مینیمم تقریبا یکسانی هستند.

با توجه به میانه در گروه انحراف ستون فقرات، متوجه میشویم که اکثر افراد دارای GS کمتری هستند و همچنین یک سری داده پرت داریم.

راه دیگر:

I=[]

for x in np.unique(data["Categories"]):#Repeat for each category

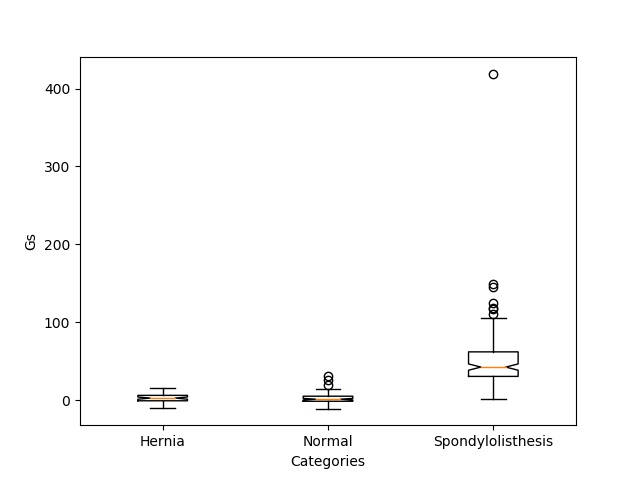
I.append(data['GS'].loc[data['Categories']==x].dropna())

plt.boxplot(I,notch=True)

plt.xticks(np.arange(1,4),np.unique(data['Categories']))

plt.xlabel('Categories')

plt.ylabel('GS')



Question 2

a)

در روش pca هدف کاهش بعد (کاهش تعداد ویژگی هاست) به طوری که بیشترین تغییر پذیری را داشتهب اشیم و اطلاعات زیادی از دست نداده باشیم. در حقیقت با این روش وابستگی میان ویژگی ها را نیز از بین میبریم (چون ابتدا ویژگی را انتخاب میکنیم که بیشترین تغییر پذیری را دارد برای انتخاب ویژگی دوم باید تغییر پذیری زیاد باید و بر ویژگی اول عمود باشد. پس واستگی از بین رفته.) هدف از استاندار کردن داده مقایسه درست تغییر پذیری ویژگی هاست. در این روش با متغیر پاسخ کاری نداریم پس تفکیک پذیری انجام نمیدیهیم در حقیقت به دنبال این نیتسیم که کدام متغیر ها توصیف خوبی از متغیر پاسخ ارائه داده اند. به دنبال این هستیم که کدام ویژگی ها انتخاب شوند به طوری که اطلاعات تهفته در دیگر متغیر ها را نیز به ما داده باشند.

پس ابتدا ستون اخر دیتا ست را کنار میگذاریم.

df=data.drop("Categories",axis=1)

df=scale(df)

df.shape

Output: 310,6

pca=PCA()

pca.fit(df)

b)

w=pca.components\_.T

w.shape

pd.DataFrame(w[:,:],index=data.columns[:-1],columns=['W1','W2','W3','W4','W5','W6'])

W1 W2 W3 W4 W5 W6

PI 0.535142 -0.002194 -0.096069 -0.102799 -0.423458 7.172898e-01

PT 0.323585 0.527545 -0.648701 -0.006441 -0.150562 -4.164920e-01

LL 0.457970 0.092875 0.152338 -0.548100 0.676770 -1.456776e-11

SS 0.445906 -0.396157 0.360313 -0.127201 -0.431500 -5.585962e-01

PR -0.143497 0.727756 0.585991 -0.174213 -0.275763 -3.446703e-12

GS 0.423978 0.162777 0.271184 0.801528 0.278852 8.303475e-12

برای تفسیر هر جهت باید به مقدار و منفی و مثبت بودن هر یک از مقادیر پرداخت. در جهت اول همه ویژگی ها به غیر از PR دارای مقدار مثبت هستند و در نتیجه تضادی بین این ویژگی ها وجود دارد یعنی هر چه این ویژگی بیشتر میشه بقیه ویژگی ها مقدار کمتری پیدا میکنند. در اولین جهت PI دارای بیشترین مقدار از نظر قدر مطلق است پس در این جهت بیشترین اطلاعات را از این ویژگی میگیریم. در جهت دوم بیشترین مقدار از نظر قدر مطلق مربوط به ویژگی PR است پس بیشترین اطلاعات در این راستا مربوط به این ویژگی است.و تضادی بین PI و SS با دیگر ویژگی هاست.

ویژگی سوم اطلاعات زیادی در مورد PT میدهد. بین دو ویژگی اول و دیگر ویژگی ها همبستگی وجود ندارد.

c)

pd.DataFrame(pca.explained\_variance\_ratio\_,index=range(1,7),columns=['Explained Variability'])

Explained Variability

1 5.409635e-01

2 1.990971e-01

3 1.268484e-01

4 7.875466e-02

5 5.433627e-02

6 6.241161e-21

نسبت هر یک از مقادیر ویژه به جمع کل مقادیر ویژه ها است. بیشترین نسبت واریانس مربوط به ویژگی اول است و بعد از آن مربوط به ویژگی دوم. میزان واراینس ویژگی اول و دوم حدود 0.3 تفاوت دارند و ویژگی دوم و سوم حدود 0.1. ویژگی سوم و چهارم حدودا 0.03. از اینجا تا ویژگی 5 تغییرات خیلی کمه.

به نظر من 3 ویژگی اول را بهتر است در نظر گرفت چون از ویژگی 4 به بعد تغییر در ورایانس کمه و اطلاعات چندانی ممکنه به ما ندهد. نمیشه گقت یک خط راست از ویژگی 4 به بعد داریم ولی اگر ویژگی 6 که خیلی نسبت واریانس ان کم است و اطلاعات چندانی به ما نمیدهد را در نظر نگیریم از ویژگی 4 به بعد یک خط راست داریم و 3 ویژگی اول کافی است.

pca.explained\_variance\_ratio\_.cumsum()

Out[41]:

array([0.54096355, 0.74006064, 0.86690906, 0.94566373, 1. ,

1. ])

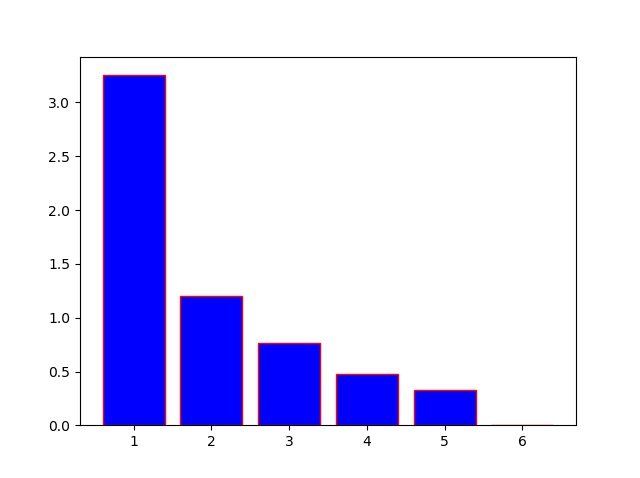
اگر میزان تجمعی واریانس ویژگی های جدید را نیز در نظر بگیریم هم به نظرم 3 ویژگی اول را بهتر است در نظر بگیریم. چون 86 درصد اطلاعات را به ما داده اگر مسئله ای داشتیم که این میزان اطلاعات از داده اولیه کم بود ویژگی 4 را هم در نظر بگیریم ولی در غیر این صورت 3 ویژگی اول کافی است.

d)

plt.figure(2)

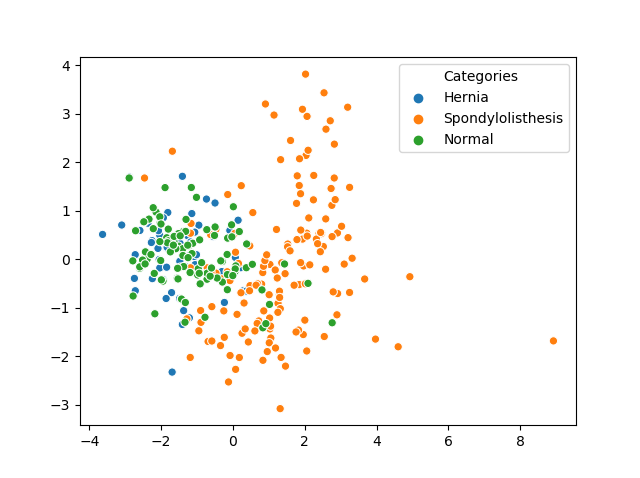
plt.bar(range(1,7),pca.explained\_variance\_,color="blue",edgecolor="Red")

بر طبق نمودار ویژگی 6 که واراینس تقریبا نزدیک به صفر داره و اطلاعات زیادی به ما نمیده پس در نظر نمیگیریم. تغییرات واریانس از ویژگی 4 به بعد کمه و شیب ان یک خط راست است پس بهتره 3 ویژگی اول را در نظر بگیریم.



e)

y=pca.transform(df)



w1 w2

PI 0.535142 -0.002194

PT 0.323585 0.527545

LL 0.457970 0.092875

SS 0.445906 -0.396157

PR -0.143497 0.727756

GS 0.423978 0.162777

این داده دارای مقدار y\_0=y[:0] مثبت زیاد است و y\_1=y[:1] آن منفی است.

چون وای صفر دارای مقدار مثبت زیاد است و و دبلیو یک دارای مقدار منفی در ویژگی پنجم است پس یا مقادیر ایکس در همه ویژگی ها جز ویژگی پنجم برای این داده زیاد بوده و در ویژگی پنجم مقدار مثبت کمی داشته و یا مقدار منفی داشته و در کل ترکیب خطی انها مثبت شده.

و اما چون دارای وای یک منفی است پس ایکس مربوط به ویژگی اول و چهارم زیاد است. در کل حالت های مختلفی بر حسب اینکه ایکس ها منفی هستند یا مثبت وجود دارد.

اگر ایکس استاندار شده مربوط به این داده را مشاهده کنیم میبینیم که

x1=data[y[:,0]>8]



این داده داری وای صفر و وای یک منفی است. اگر تمام مقادیر ایکس این داده مثبت باشند چون وای صفر منفی شده پس مقدار ایکس در ویژگی پنجم خیلی زیاد است و ایکس بقیه کم است. اما در این صورت وای یک نمیتواند منفی باشد.

پس حتما ایکس منفی داشته ایم.

بله داده ای که مقدار وای صفر بیشتر از 8 و وای یک منفی دارد داده ای پرت است.

داده ای که با مربع نشان داده شده است هم از دیگر نقاط ابی رنگ دور است و در این کلاس داده ای پرت است.

f)

I)

داده که وای صفر بیشتر از 8 و وای یک منفی دارد متعلق به کلاس لغزش ستون فقرات است.

II)

data[y[:,0]>8]



Mean-3\*std, Mean+3\*std

PI: (8.87,112.122)

PT: (-12.43, 47.519)

LL: ) -3.64, 107.503)

SS: (2.74, 83.15)

PR: (78.03, 157.80)

GS: (-86.19, 138.79)

اگر بدون در نظر گرفتن کلاس ها بازه سه برابر و منفی سه برابر انحراف معیار را در نظر بگیریم این داده بر اساس ویژگی های PI,SS,GS داده ای پرت است.

اما با در نظر گرفتن کلاس لغزش ستون فقرات داریم:

PI: (26.18, 116.84)

PT: (-13.77, 55.266)

LL: (14.91, 113.30)

SS: (13.809, 87.72)

PR: (67.7, 161.25)

GS: (-68.42, 172.22)