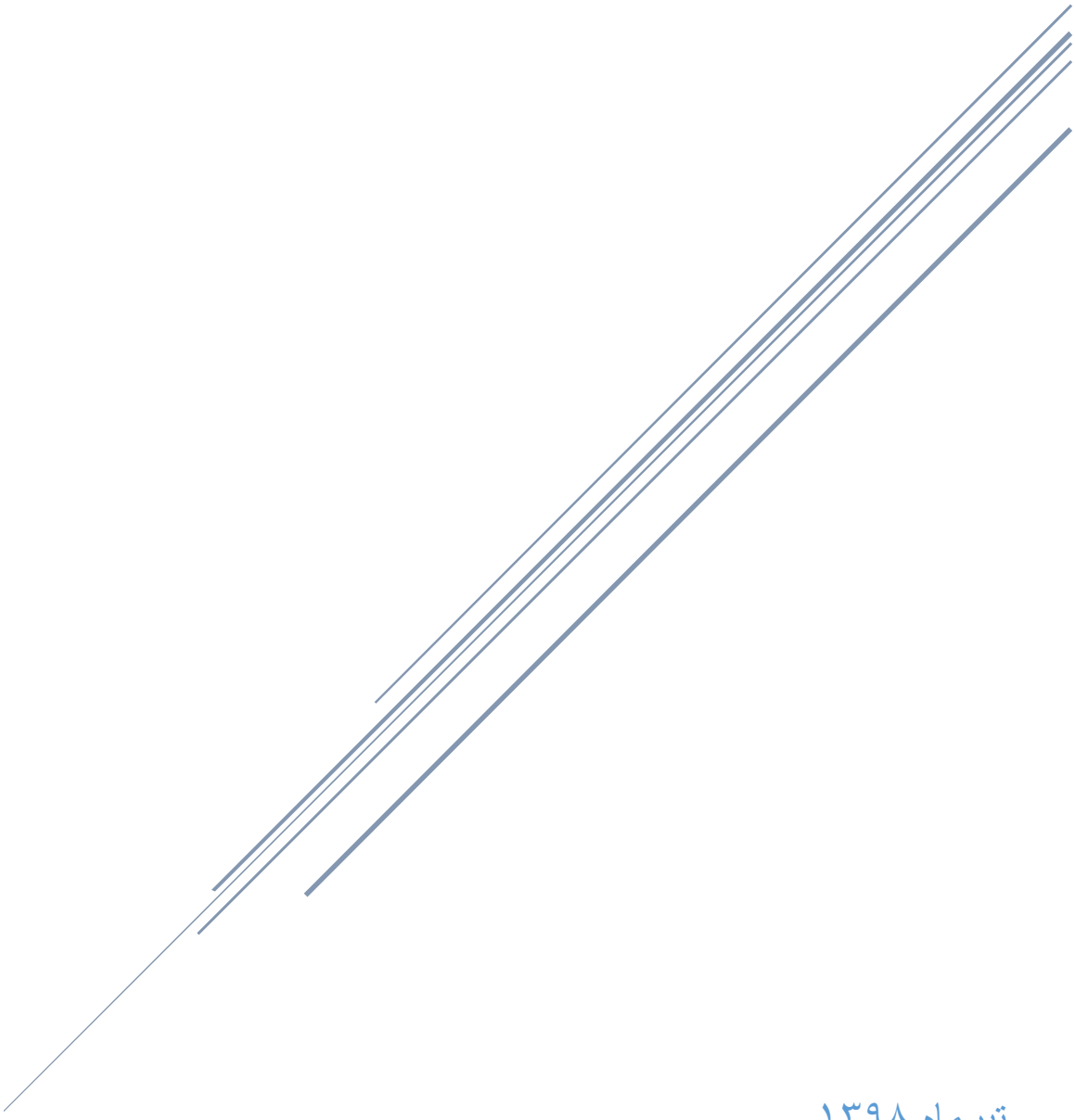
گزارش کار پروژه طراحی دارو

علیرضا دوست محمدی



تیرماه ۱۳۹۸

دانشگاه تربیت مدرس

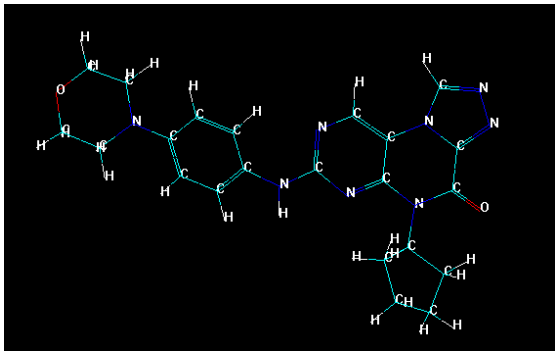
مقدمه:

داروها مولکول های ریزی با وزن کمتر از ۵۰۰ دالتون هستند که جهت جلوگیری از عملکرد مخرب ها مورد استفاده قرار میگرند. فضای مولکول های شیمیایی[۱] با حداکثر ۱۶ اتم برای هر مولکول برابر $۱۰^{۶۰}$ مولکول است حال آنکه تعداد مولکول های شیمیایی سنتز شده در هر سال در آزمایشگاه های دنیا در مجموع حدود یک ملیون مولکول می باشد. بنابراین سنتز مولکول ها نمی تواند راهکار مناسبی جهت ایجاد تمامی حالات فضای مولکول های شیمیایی باشد. ریموند[۲] جهت مدل کردن این فضای شیمیایی از روش غربالگری مجازی[۳] استفاده کرد و به این شکل فضای شیمیایی را مدل کرد. روش های طراحی دارو به دو دسته مبتنی بر لیگاند و مبتنی بر ساختار تقسیم می شوند. در روش مبتنی بر لیگاند به طور مستقیم از ساختار مولکول استفاده نمی کنیم و به طور کلی به شامل دو زیر دسته LBVS و QSAR می باشد. در روش QSAR تعداد مولکول های مورد استفاده جهت شبیه سازی اندک است و هدف پیش بینی دقیق فعالیت مولکول های شیمیایی می باشد. در این پروژه از روش QSAR استفاده شده است.

در پروژه انجام شده ابتدا مولکول های شیمیایی مقاله انتخابی [۱] در محیط کامپیوتر رسم شده است و پس از بهینه سازی، ویژگی های مولکول ها استخراج و پس از نرمال سازی با استفاده از روش های MLR و PCR و PLS و شبکه عصبی و ترکیب الگوریتم ژنتیک با PLS و MLR و شبکه عصبی مدل سازی انجام شده است. مراحل انجام شده و نتایج حاصل به شرح زیر است:

۱. رسم مولکول های شیمیایی

مقاله مورد استفاده شامل ۳۴ مولکول شبیه به هم[۴] با مقادیر IC۵۰ مشخص می باشد. بازه میزان فعالیت مولکول ها ۰,۱۶ تا ۴۱,۶۸ می باشد. جهت استخراج ویژگی های مولکول ها و مدل سازی مبتنی بر آن ابتدا هر کدام از مولکول ها در نرم افزار HyperChem رسم شده و سپس در دو مرحله بهینه سازی شده است. در ابتدا بر اساس قوانین کوآنتوم مکانیک و با استفاده از متود MM+ بهینه سازی انجام شده است. با توجه به آنکه این نوع بهینه سازی بر اساس ساختار خطی مولکول می باشد و با ساختار طبیعی مولکول متفاوت است، این نوع بهینه سازی در بازه زمانی کوتاهی در حد ۵ ثانیه انجام شده و سپس با استفاده از روش های تجربی و متود AM۱ بهینه سازی مولکول ها ادامه پیدا کرده است.



شکل ۱: نمونه ای از مولکول های بهینه شده در نرم افزار HyperChem

---

[۱] Chemical Space
[۲] Reymond
[۳] Virtual Screening
[۴] Homolog

۲. استخراج ویژگی های مولکول ها

فایل های خروجی hin. حاصل از رسم و بهینه سازی مولکول ها به عنوان ورودی نرم افزار Dragon جهت استخراج ویژگی های مختلف مولکول ها استفاده شده است. نرم افزار Dragon قادر است حداکثر ۳۲۲۴ ویژگی از مولکول های شیمیایی استخراج کند. ویژگی های استخراج شده در قالب چهار دسته صفر بعدی، یک بعدی، دو بعدی و سه بعدی قابل تقسیم است. ویژگی های یک بعدی ویژگی هایی هستند که جهت محاسبه آنها فرمول کلی مولکول کفایت می کند حال آنکه جهت محاسبه ویژگی های یک بعدی نیازمند ساختار کلی مولکول و در دو بعدی نیازمند ساختار دو بعدی مولکول و در سه بعدی نیازمند ساختار سه بعدی مولکول هستیم. ویژگی های استخراج شده توسط Dragon برای این پروژه ۱۳۵۸ویژگی می باشد و جهت مدل سازی در نرم افزار Matlab استفاده شده است.

۳. مدل سازی مبتنی بر QSAR:

۳,۱. بررسی معیارعامل تورم واریانس جدول ویژگی ها[۵]

این معیار نشاندهنده میزان وابستگی ویژگی ها به یکدیگر می باشد و از رابطه زیر محاسبه می شود:

$$ViF = \frac{۱}{۱ - R^{۲}{}_{m}}$$

$$R^{۲}{}_{m} = \frac{\sum_{i=۱}^{m} R^{۲}{}_{i}}{m} \quad ; m = number\ of\ Featuures$$

اگر مقدار VIF بیشتر از ۲ باشد بدین معناست که وابستگی ویژگی ها بسیار زیاد است و ۵۰ درصد ویژگی ها را می توان از بقیه ویژگی ها بدست آورد و نیازمند آن است که از میان ویژگی هایی که وابستگی زیادی به یکدیگر دارند (به طور مثال وابستگی بیشتر از ۰,۹)، ویژگی ای که وابستگی کمتری نسبت به دیگر ویژگی ها دارد را انتخاب و بقیه را حذف کنیم. مقدار VIF ماتریس داده های اینجانب برابر ۱,۰۸۸۸ شد و نیاز به کاهش ابعاد ماتریس نمی باشد.

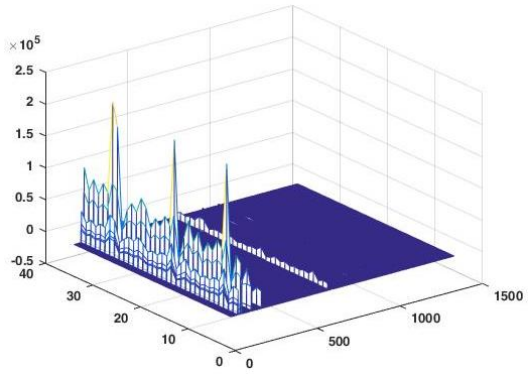
۳,۲. نرمال سازی داده ها[۶]

از آنجا که رنج داده های هر ستون از ماتریس داده ها با یکدیگر متفاوت است و این موضوع در مدل سازی بسیار تاثیر گذار است نیازمند نرمال سازی داده ها هستیم تا رنج داده ها به طور تقریبی در هر ستون یکسان شود. برای این کار از چهار روش مختلف بهره میبریم که الگوریتم هر یک به شرح زیر است:

---

[۵] Variance Inflation Factor(VIF)

[۶] Auto Scaling



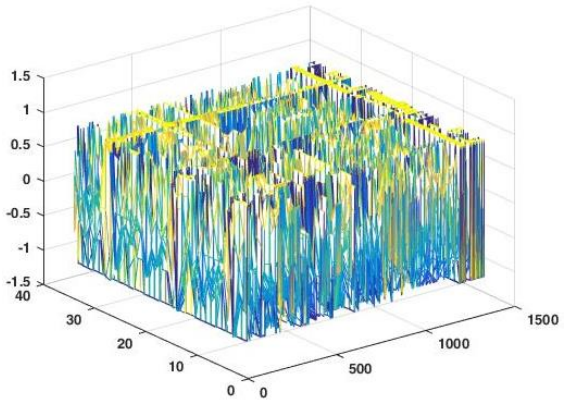
نمودار ۱: توزیع مقادیر ویژگی ها پیش از نرمال سازی

۳,۲,۱. تغییر رنج داده ها به بازه [+۱,-۱]:

```
for col=١:size(X,٢)
    maxVal=max(X(:,col));
    minVal=min(X(:,col));

    sum=maxVal+minVal;
    mine=maxVal-minVal;
    X١(:,col)=((X(:,col)*٢)/mine)-(sum/mine);
end
```

توزیع مقادیر ویژگی ها پس از اجرای این متود به شرح زیر است:

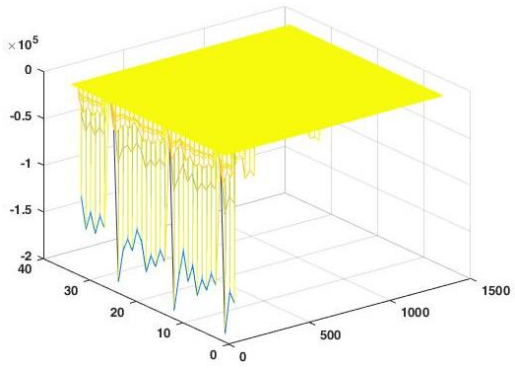


نمودار۲: توزیع مقادیر ویژگی ها با نرمال سازی متود ۱

۳،۲،۲. نرمال سازی بر اساس بیشینه مقدار هر ستون:

```
for col=١:size(X,٢)
    maxVal=max(X(:,col));

   X(:,col)=X(:,col)/maxVal;
end
```

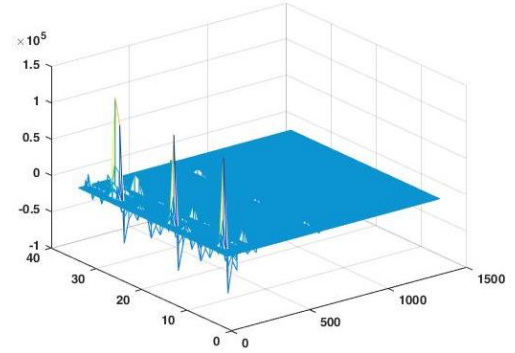


نمودار۳: توزیع مقادیر ویژگی ها با نرمال سازی متود ۲

۳،۲،۳. نرمال سازی بر اساس مقدار میانه هر ستون:

```
for col=١:size(X,٢)
   meanVal=mean(X(:,col));

   X(:,col)=X(:,col)-meanVal;
end
```

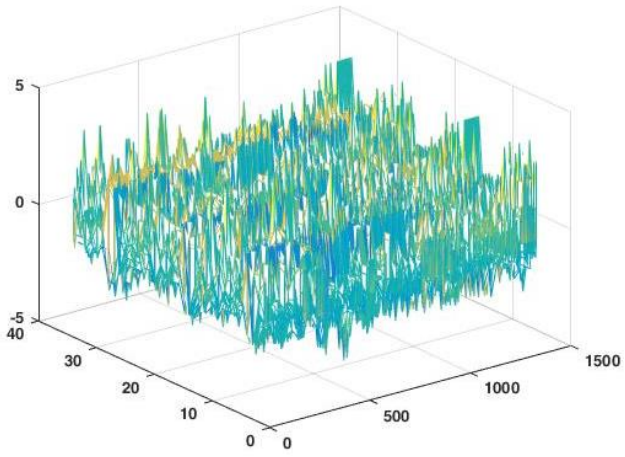


نمودار۴: توزیع مقادیر ویژگی ها با نرمال سازی متود ۳

۳،۲،۴. نرمال سازی بر اساس مقدار میانه و انحراف معیار هر ستون:

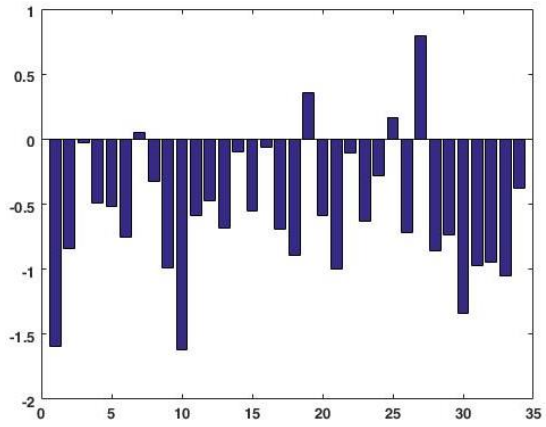
```
for col=١:size(X,٢)
    meanVal=mean(X(:,col));
    stdVal=std(X(:,col));

    X(:,col)=(X(:,col)-meanVal)/stdVal;
end
```



نمودار۵: توزیع مقادیر ویژگی ها با نرمال سازی متود ۴

پس از انجام نرمال سازی با استفاده از داده های نرمال شده به واسطه متود ۱ و ۴ به مدل سازی می پردازیم. پیش از آن باید از IC۵۰ های گزارش شده در مقاله log – در پایه ده بگیریم.(نمودار ۶)



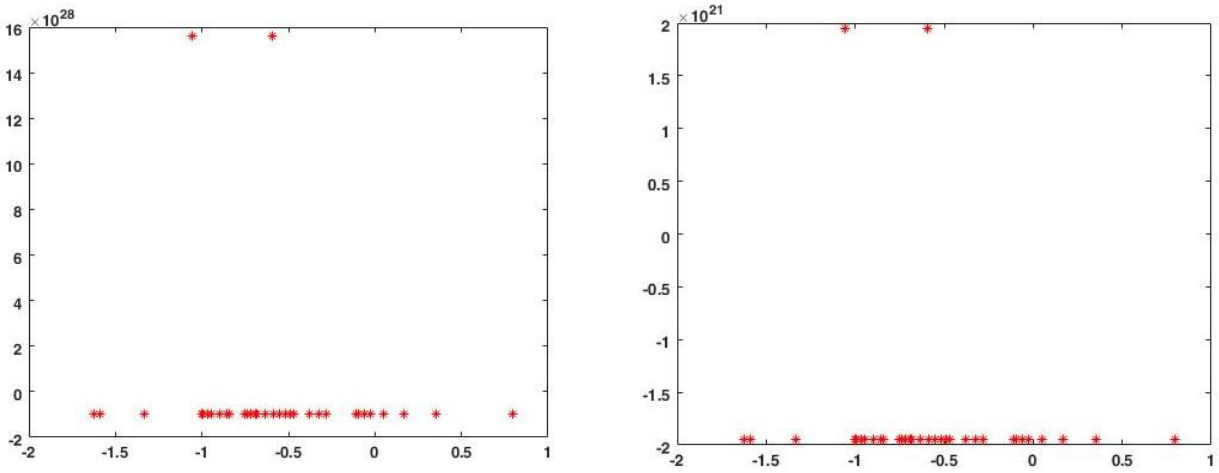
نمودار۶: $-log$ مقدار IC۵۰ ها

۳،۳. مدل سازی

پس از نرمال سازی داده ها مدل های MLR، PLS،PCR، GA-PLS،NN، GA-MLR و GA-NN بر روی داده های حاصل از نرمال سازی به واسطه متود ۱ و متود ۴ اجرا و نتایج زیر حاصل شد:

۳،۳،۱. Multiple Linear Regression (MLR):

با توجه به آنکه ابعاد ماتریس دیتاها ۱۳۵۶ * ۳۴ است، اگر مدل MLR را با کل ماتریس ایجاد کنیم احتمالا مدل حاصل به علت وابستگی ویژگی ها به یکدیگر مدل خوبی نخواهد بود و $RMSE$,R^2 خوبی نخواهد داشت. بنابراین با استفاده از روش Stepwise ابتدا ابعاد ماتریس را کاهش میدهیم و سپس مدل سازی انجام می دهیم. جهت کاهش ابعاد ماتریس از دستور stepwisefit() استفاده شده است.



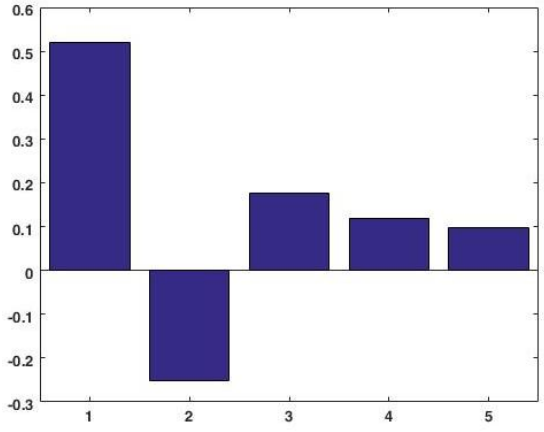
نمودار۷: مدل MLR بر روی تمام ویژگی های داده های نرمال شده با متود ۱ و ۴

جدول ۱: کاهش ابعاد با استفاده از روش stepwise

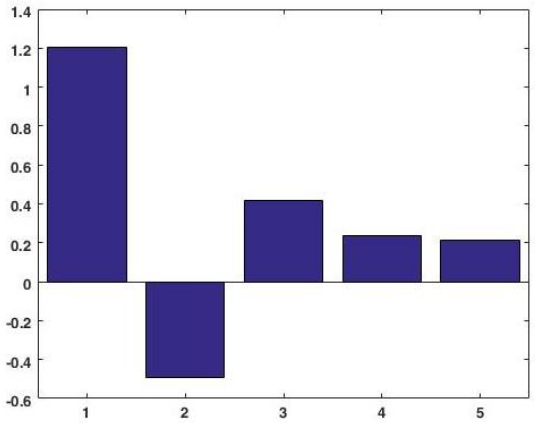
ماتریس ورودی	ستون های انتخابی	R^2	$RMSE$
نرمال شده در رنج [-۱,+۱][۷]	۹۲۹,۹۵۸,۲۹۴,۷۹۹,۸۴۱	۰,۷۸۱۵	۰,۲۳۸۴
نرمال شده بر اساس مقدار میانه و انحراف معیار هر ستون[۸]	۹۲۹,۹۵۸,۲۹۴,۷۹۹,۸۴۱	۰,۷۸۱۵	۰,۲۳۸۴

[۷] در طول گزارش کار به جای این واژه از واژه نرمال شده با متود ۱ استفاده خواهد شد.
[۸] در طول گزارش کار به جای این واژه از واژه نرمال شده با متود ٤ استفاده خواهد شد.

ضرایب حاصل از مدل سازی stepwise - MLR به صورت زیر است: (نمودار ۸ و ۹)

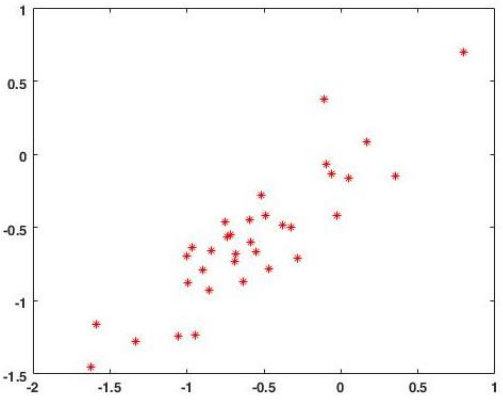


نمودار۸: ضرایب مدل ۲

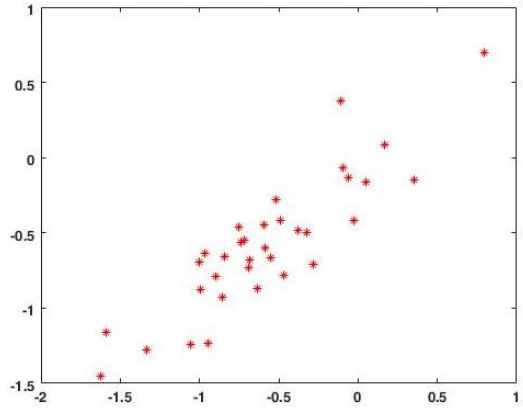


نمودار۹: ضرایب مدل ۱

مقایسه مقدار IC_{50} پیش بینی شده توسط مدل و مقدار دقیق آن در نمودارهای زیر قابل ملاحظه است: (نمودار ۱۰ و ۱۱)



نمودار۱۰: مقایسه Y , $\hat{Y}$ مدل ۲



نمودار۱۱: مقایسه Y , $\hat{Y}$ مدل ۱

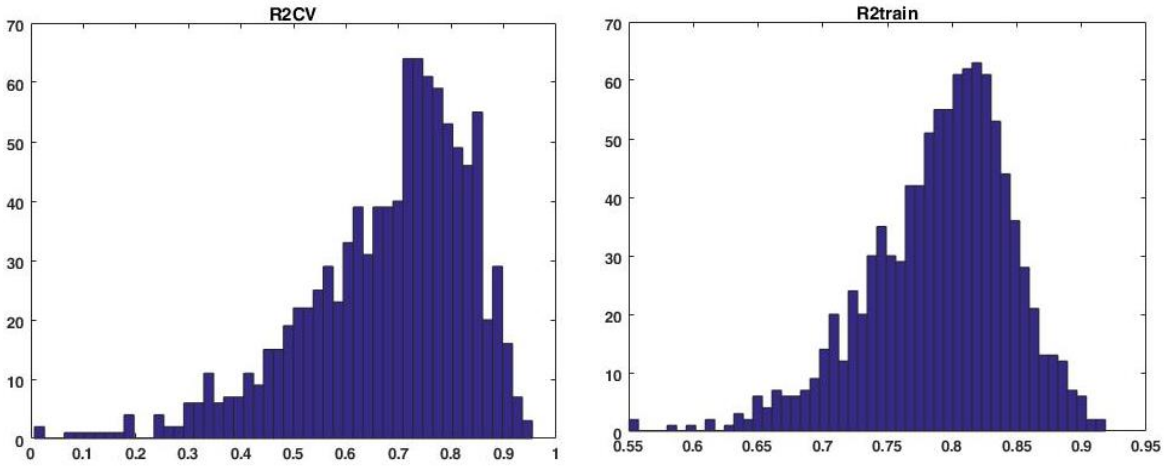
حال مدل ساخته شده را با استفاده از روش LOO[۹] و MMC[۱۰] ارزیابی می کنیم: (جدول شماره ۲)

جدول ۲: ارزیابی مدل MLR ساخته شده

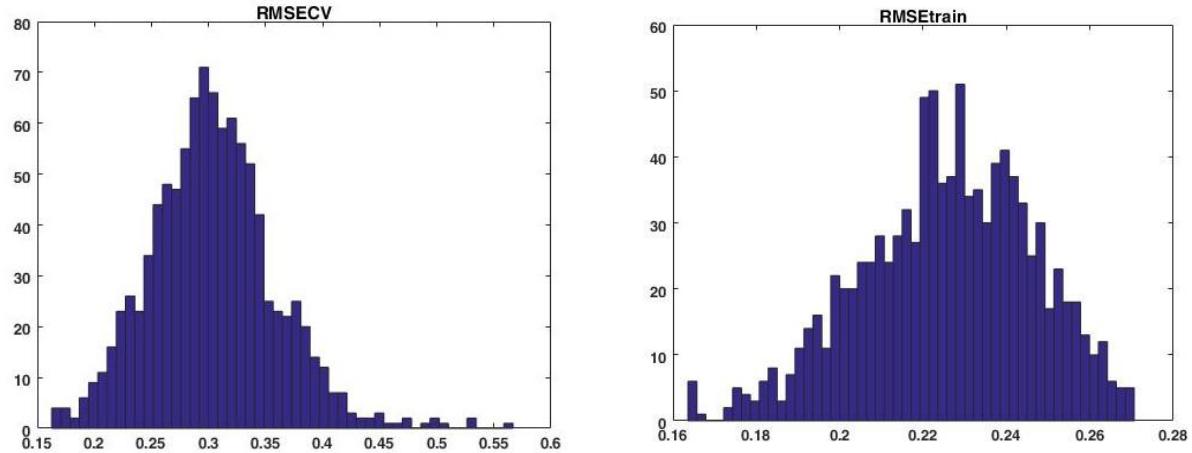
	Train		Test- MCCV		Test-LOO	
ماتریس ورودی	R^2C	$RMSEC$	R^2V-MCC	$RMSEV-MCC$	R^2V-LOO	$RMSEV-LOO$
نرمال شده با متود ۱	۰,۷۸۱۵	۰,۲۳۸۴	۰,۶۸۳۷	۰,۳۰۳۹	۰,۶۸۲۵	۰,۲۹
نرمال شده با متود ۴	۰,۷۸۱۵	۰,۲۳۸۴	۰,۶۸۴۳	۰,۳۰۵۶	۰,۶۸۲۵	۰,۲۹

[۹] Leave One Out
[۱۰] Monte Carlo

در نمودارهای زیر مقایسه میان توزیع R^2CV و $RMSE-CV$ با توزیع R^2train و $RMSE-train$ در طول ۱۰۰۰ بار تکرار مونت کارلو قابل مشاهده است.



نمودار۱۲: مقایسه میان توزیع R^2CV , R^2train مدل ۱

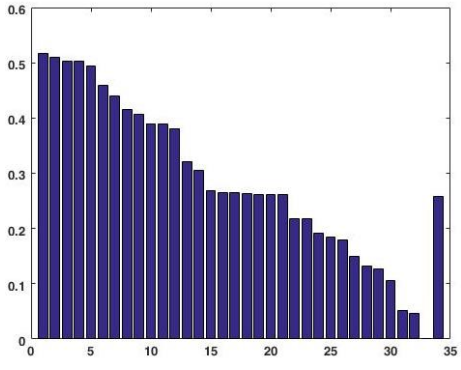


نمودار۱۳: مقایسه میان توزیع $RMSECV$, $RMSEtrain$ مدل ۱

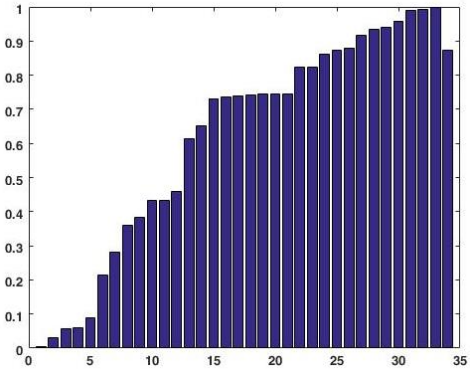
۳،۳،۲. Principal Component Regression (PCR):

از الگوریتم های یاد شده جهت مدل سازی داده ها، ورودی دو مدل PCR و NN ماتریس PCA است. اگر با PCA یک مدل خطی بسازیم حاصل PCR و اگر مدل غیر خطی بسازیم حاصل NN می شود.

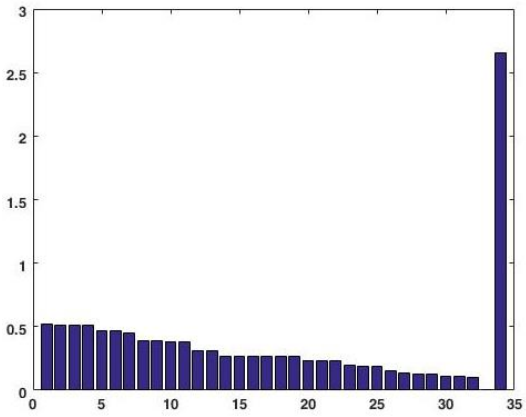
جهت یافتن بهترین تعداد کامپوننت PCA جهت ساخت مدل PCR مقادیر R^2 , $RMSE$ به ازای تعداد کامپوننت های مختلف را بررسی می کنیم: (نمودار ۱۴و ۱۵ و ۱۶ و ۱۷)



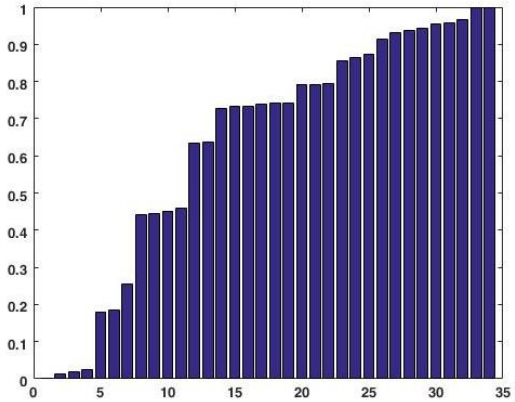
نمودار ۱۴: مقدار RMSE تعداد کامپوننت های مختلف بر روی داده های متود ۱



نمودار ۱۵: مقدار R^2 تعداد کامپوننت های مختلف بر روی داده های متود ۱

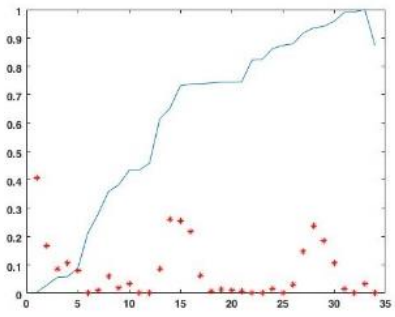


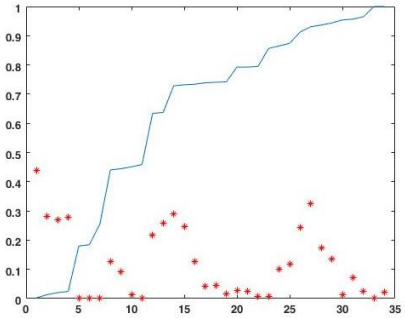
نمودار ۱۶: مقدار RMSE تعداد کامپوننت های مختلف بر روی داده های متود ۴



نمودار ۱۷: مقدار R^2 تعداد کامپوننت های مختلف بر روی داده های متود ۴

از نمودارهای ۱۴ تا ۱۷ متوجه می شویم تعداد کامپوننت مناسب برای داده های متود ۱ برابر ۲۷ و برای داده های متود ۴ برابر ۲۶ است. اما این انتخاب در صورتی اعتبار پیدا می کند که نتایج حاصل از بررسی $R^2\ Validation\ ,RMSE\ Validation$ نیز این تعداد کامپوننت را انتخاب نشان دهد.

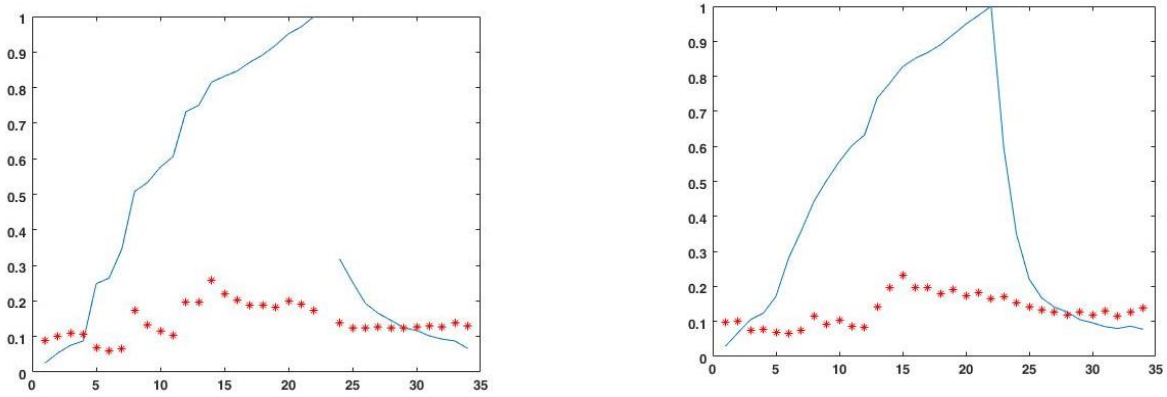




نمودار۱۷: مقایسه مقدار $R^2\ train\ ,R^2\ test$ با استفاده از روش LOO - شکل سمت چپ داده های متود ۴ و شکل سمت راست داده های متود ۱

آنطور که از نمودار ۱۷ مشخص می شود تعداد کامپوننت برای بهترین حالت مدل سازی با PCR برابر ۱۴ کامپوننت برای داده های هر دو متود ۱ و ۴ است و همانطور که ملاحظه می شود مقدار R^2 برای هر دو حالت test , train بسیار ضعیف است. این موضوع تا مقداری قابل پیش بینی بود زیرا ممکن است در مدل PCR ساخته شده میان PC ها و result وابستگی زیادی وجود نداشته باشد.

اجرای ارزیابی PCR با تعداد کامپوننت های مختلف به روش مونت کارلو نیز جواب مشابهی به ما می دهد.

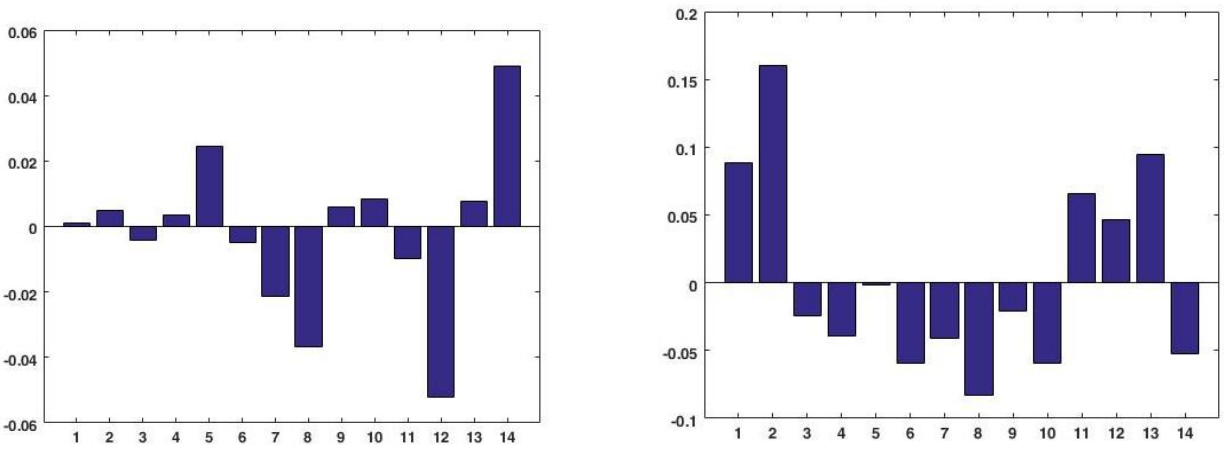


نمودار۱۸: مقایسه مقدار $R^2\ train$, $R^2\ test$ با استفاده از روش MCC- شکل سمت چپ داده های متود ۴ و شکل سمت راست داده های متود ۱

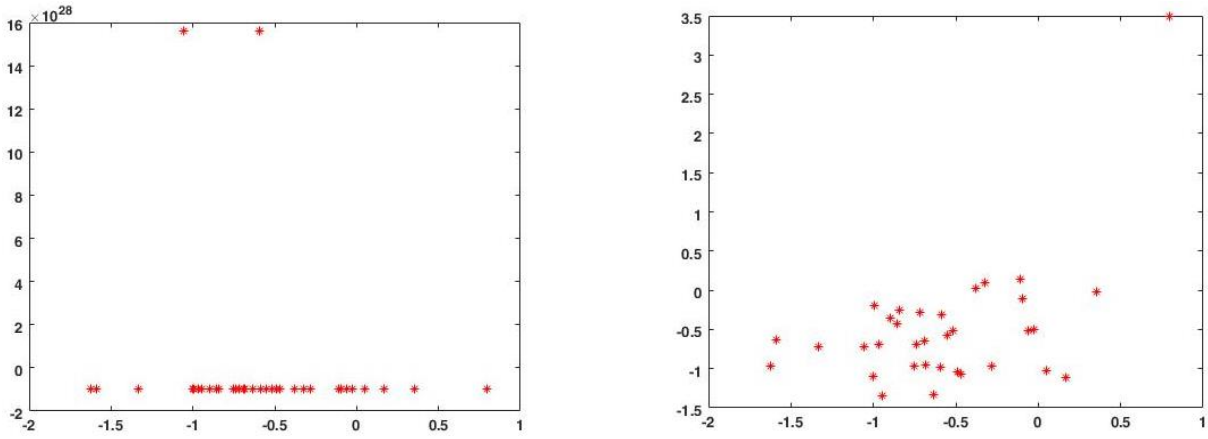
ضرایب و مدل های ساخته شده در نمودارهای زیر قابل شهود است:

جدول ۳: ارزیابی مدل های PCR ساخته شده

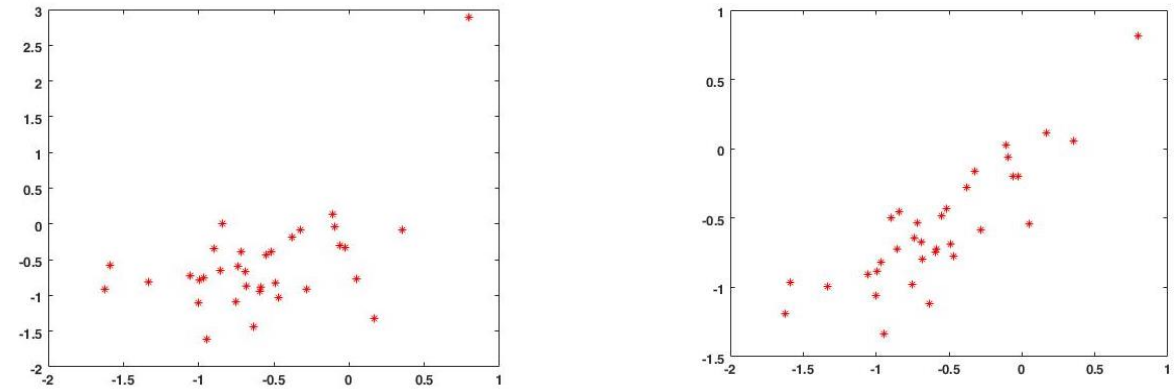
	Components	Train		Test- MCCV		Test-LOO	
ماتریس ورودی		R^2C	$RMSEC$	R^2V-MCC	$RMSEV-MCC$	R^2V-LOO	$RMSEV-LOO$
نرمال شده با متود ۱	۱۴	۰,۶۵۲۲	۰,۳۰۵۳	۰,۱۹۶۶	۲,۰۳۱۲	۰,۲۵۹۶	۰,۷۰۱۲
نرمال شده با متود ۴	۱۴	۰,۹۱۲۹	۰,۱۵۲۷	۰,۲۴۷۶	۱,۶۰۷۲	۰,۲۸۲۹	۰,۶۲۳۹



نمودار ۱۹: ضرایب مدل داده های نرمال شده با متود ۱ و متود ۴ – نمودار سمت راست ضرایب متود ۱ و نمودار سمت چپ ضرایب متود ۴ است



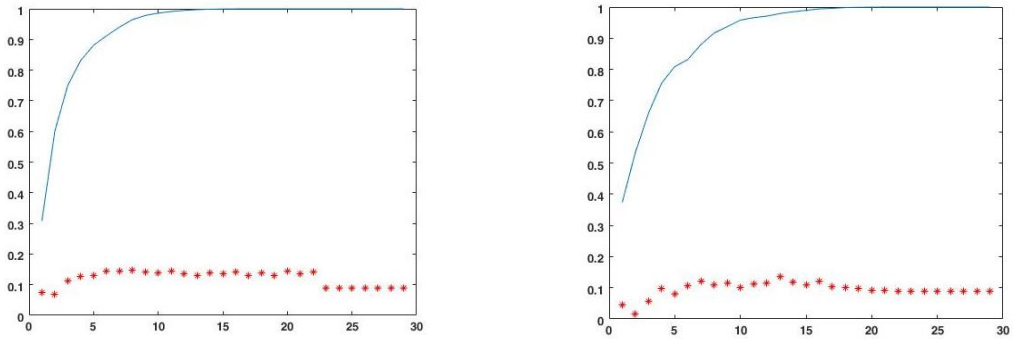
نمودار ۲۰: مدل PCR ماتریس نرمال متود ۱ در حالت های train و test – نمودار سمت راست حالت train و نمودار سمت چپ حالت test را نمایش می دهد.



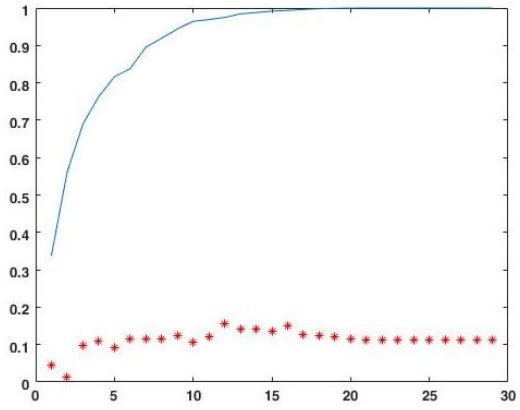
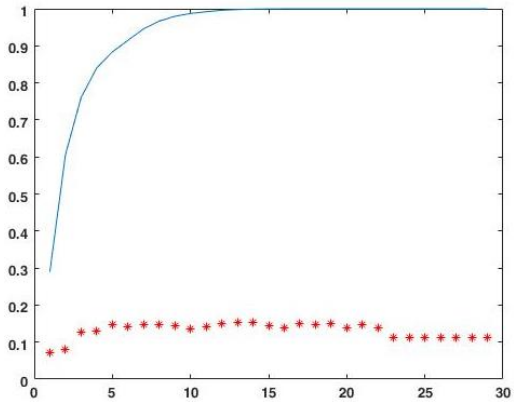
نمودار ۲۱: مدل PCR ماتریس نرمال متود ۴ در حالت های train و test – نمودار سمت راست حالت train و نمودار سمت چپ حالت test را نمایش می دهد.

۳,۳,۳ Partial least Squares Regression(PLS):

جهت یافتن تعداد کامپوننت مناسب برای PLS مقدار R^2 حاصل از Calibrate و Validation را با هم مقایسه می کنیم.



نمودار ۲۲: مقدار R^2 مدل PLS به ازای تعداد کامپوننت های مختلف داده های نرمال ۱ – نمودار سمت راست به روش LOO و نمودار سمت چپ به روش MCC ارزیابی است.

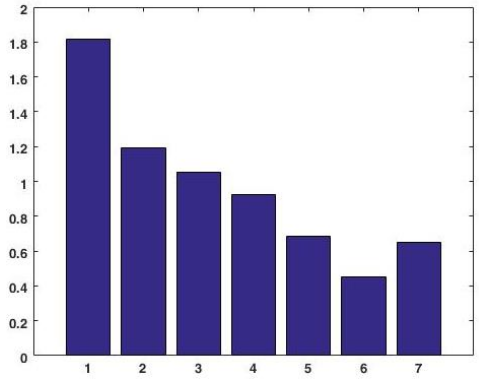
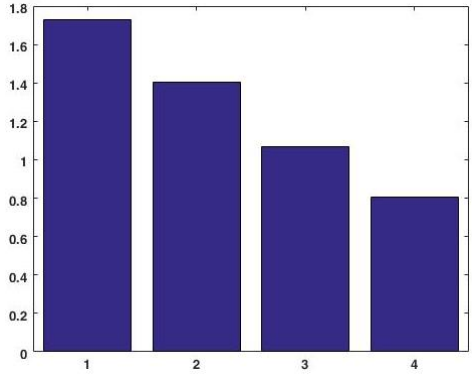
نمودار ۲۳: مقدار R^2 مدل PLS به ازای تعداد کامپونت های مختلف داده های نرمال ۴ – نمودار سمت راست به روش LOO و نمودار سمت چپ به روش MCC ارزیابی است.

همانطور که از نمودار ۲۲ و ۲۳ قابل مشاهده است، بهترین تعداد کامپونت جهت مدل سازی PLS بر روی داده های نرمال ۱ برابر ۷و برای داده های نرمال ۴ برابر ۴ کامپونت است. انتخاب ۱۳ کامپونت برای داده های نرمال ۴ موجب overfit شدن مدل خواهد شد به همین دلیل تنها به انتخاب ۸ کامپونت کفایت کردیم.

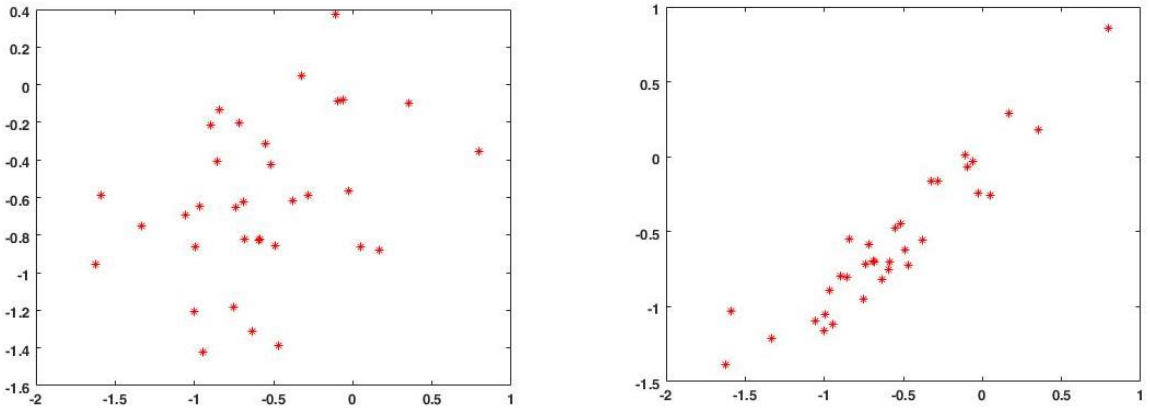
ضرایب و مدل های ساخته شده به شرح زیر هستند:

جدول ۴: ارزیابی مدل های PLS ساخته شده

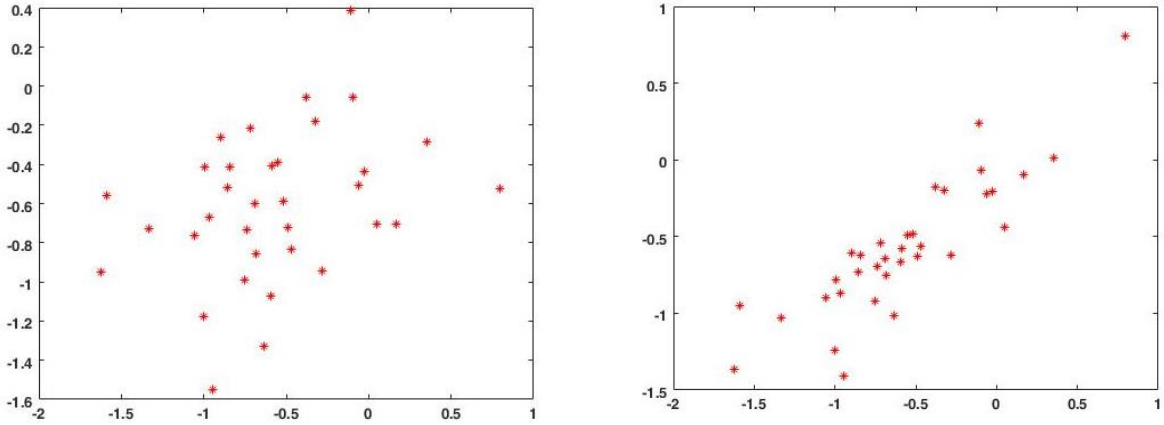
	Components	Train		Test- MCCV		Test-LOO	
ماتریس ورودی		R^2C	$RMSEC$	R^2V-MCC	$RMSEV-MCC$	R^2V-LOO	$RMSEV-LOO$
نرمال شده با متود ۱	۷	۰,۹۴۱۲	۰,۲۴۸۵	۰,۱۴۳۰	۰,۵۸۴۸	۰,۱۲۱۱	۰,۵۳۸۱
نرمال شده با متود ۴	۴	۰,۸۳۹۱	۰,۱۹۸۸	۰,۱۲۹۴	۰,۵۵۷۳	۰,۲۸۲۹	۰,۱۰۸۴

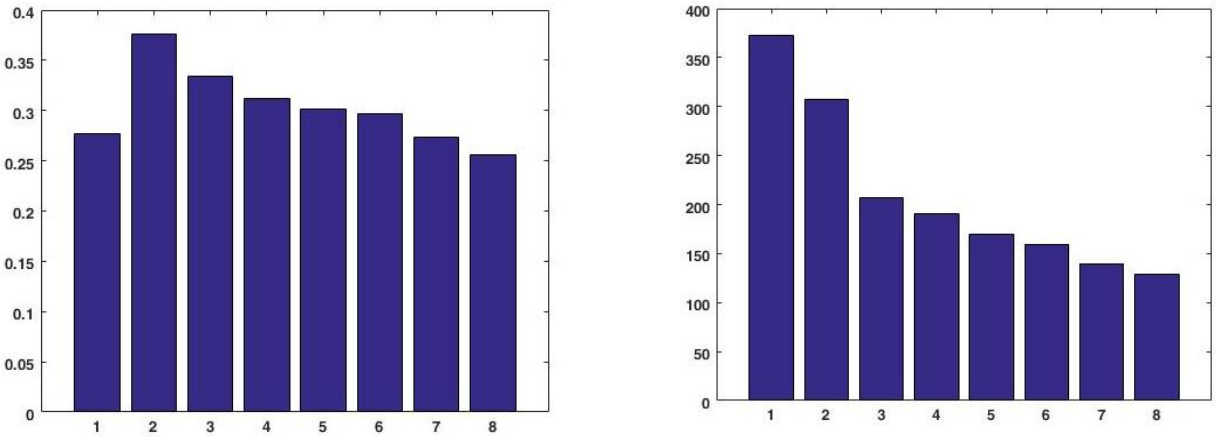
نمودار ۲۴: ضرایب مدل داده های نرمال شده با متود ۱ و متود ۴ – نمودار سمت راست ضرایب متود ۱ و نمودار سمت چپ ضرایب متود ۴ است



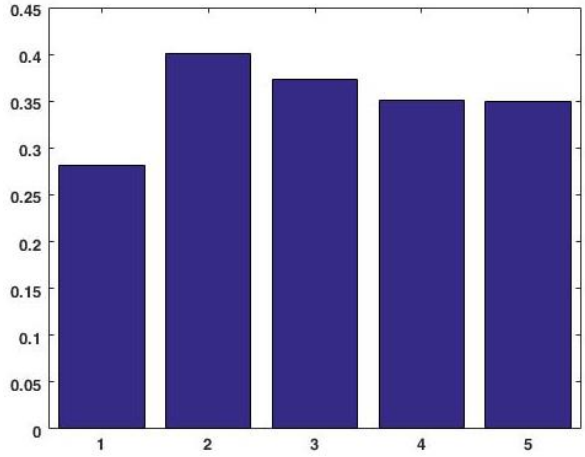
نمودار۲۵: مدل PLS ماتریس نرمال متود ۱ در حالت های train و test – نمودار سمت راست حالت train و نمودار سمت چپ حالت test را نمایش می دهد.

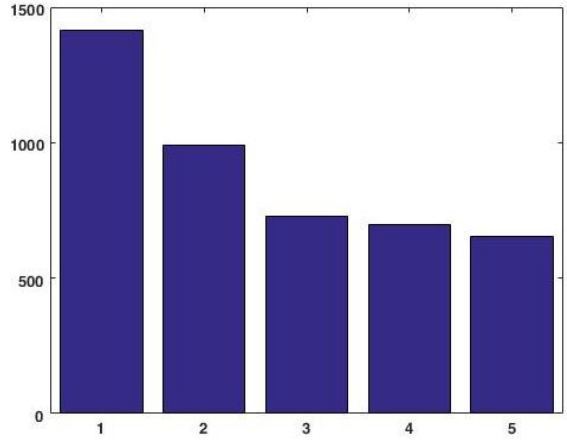


نمودار۲۶: مدل PLS ماتریس نرمال متود ۴ در حالت های train و test – نمودار سمت راست حالت train و نمودار سمت چپ حالت test را نمایش می دهد.



نمودار۲۷: میانگین مربع خطاها در X و Y داده های نرمال ۱





نمودار۲۸: میانگین مربع خطاها در X و Y داده های نرمال ۴

حال به مقایسه هر سه مدل PCR، PLC و Stepwise MLR می پردازیم:

جدول ۵: مقایسه مدل های StepWise MLR و PCR و PLS

ماتریس ورودی	Model	Components/ Features	Train		Test- MCCV		Test-LOO	
			$RMSEC$	R^2C	$RMSEV - MCC$	$R^2V - MCC$	$RMSEV - LOO$	$R^2V - LOO$
نرمال شده با متود ۱	StepWiseMLR	۹۲۹,۹۵۸,۲۹۴,۷۹۹,۸۴۱	۰,۲۳۸۴	۰,۷۸۱۵	۰,۳۰۳۹	۰,۶۸۳۷	۰,۲۹	۰,۶۸۲۵
	PCR	۱۴	۰,۳۰۵۳	۰,۶۵۲۲	۰,۱۹۶۶	۲,۰۳۱۲	۰,۷۰۱۲	۰,۲۵۹۶
	PLS	۷	۰,۲۴۸۵	۰,۹۴۱۲	۰,۱۴۳۰	۰,۵۸۴۸	۰,۵۳۸۱	۰,۱۲۱۱
نرمال شده با متود ۴	StepWiseMLR	۹۲۹,۹۵۸,۲۹۴,۷۹۹,۸۴۱	۰,۲۳۸۴	۰,۷۸۱۵	۰,۳۰۵۶	۰,۶۸۴۳	۰,۲۹	۰,۶۸۲۵
	PCR	۱۴	۰,۱۵۲۷	۰,۹۱۲۹	۰,۲۴۷۶	۱,۶۰۷۲	۰,۶۲۳۹	۰,۲۸۲۹
	PLS	۴	۰,۱۹۸۸	۰,۸۳۹۱	۰,۱۲۹۴	۰,۵۵۷۳	۰,۱۰۸۴	۰,۲۸۲۹

همانطور که ملاحظه می شود، مدل MLR نتایج بهتری نسبت به دیگر مدل ها ارائه داده است.

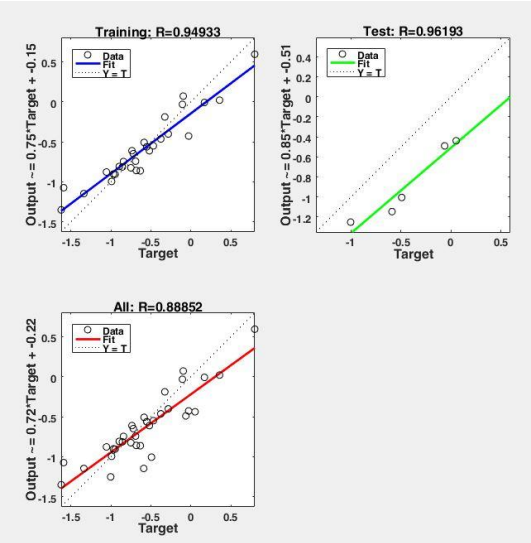
۳,۳,۴. Neural Network:

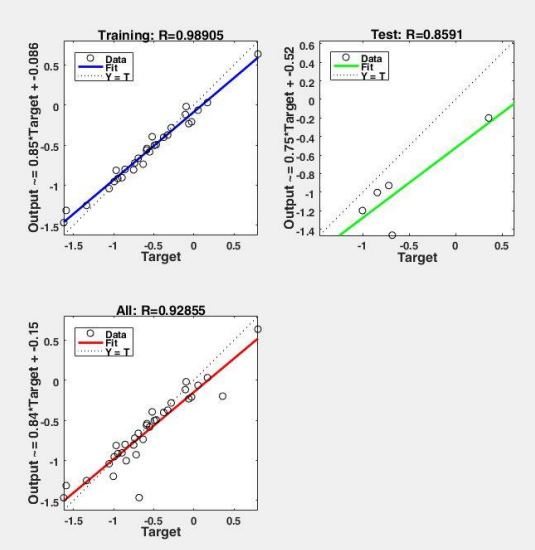
در این پروژه از شبکه عصبی لایه ای با چهار لایه پنهان[۱۱] استفاده شده است. به طور معمول انتظار داریم نتیجه حاصل از شبکه عصبی از مدل های MLR و PCR و PLS بهتر باشد و این موضوع را تحقیق می کنیم. ورودی شبکه عصبی همانند PCR، PCA است با این تفاوت که در PCR با PCA ها یک مدل خطی می ساختیم اما درشبکه عصبی یک مدل غیر خطی ایجاد می کنیم.

[۱۱] Hidden Layer

جدول ۶: ارزیابی مدل های شبکه عصبی ساخته شده

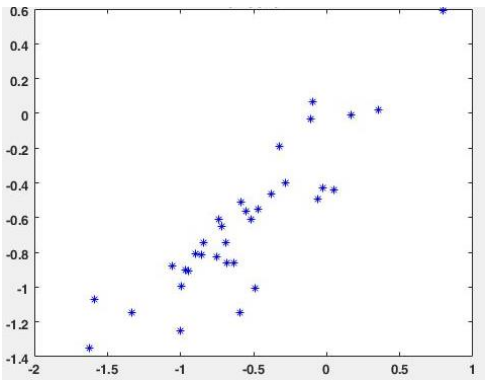
ماتریس ورودی	PCA Components	R^2C	R^2V
نرمال شده با متود ۱	۳۰	۰,۸۶۲۲۱	۰,۷۳۸۰
نرمال شده با متود ۴	۲۳	۰,۷۸۹۵	۰,۹۲۵۳

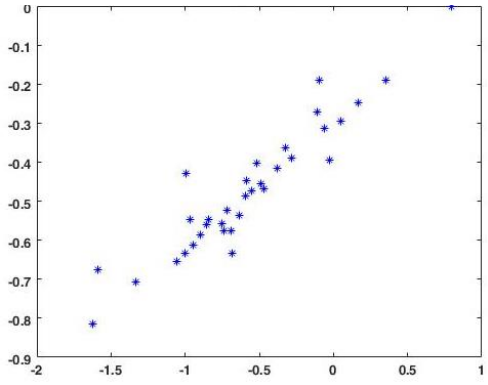




نمودار۲۹: مقدار خطای R در حالت test , train داده های نرمال شده با متود ۱ و ۲

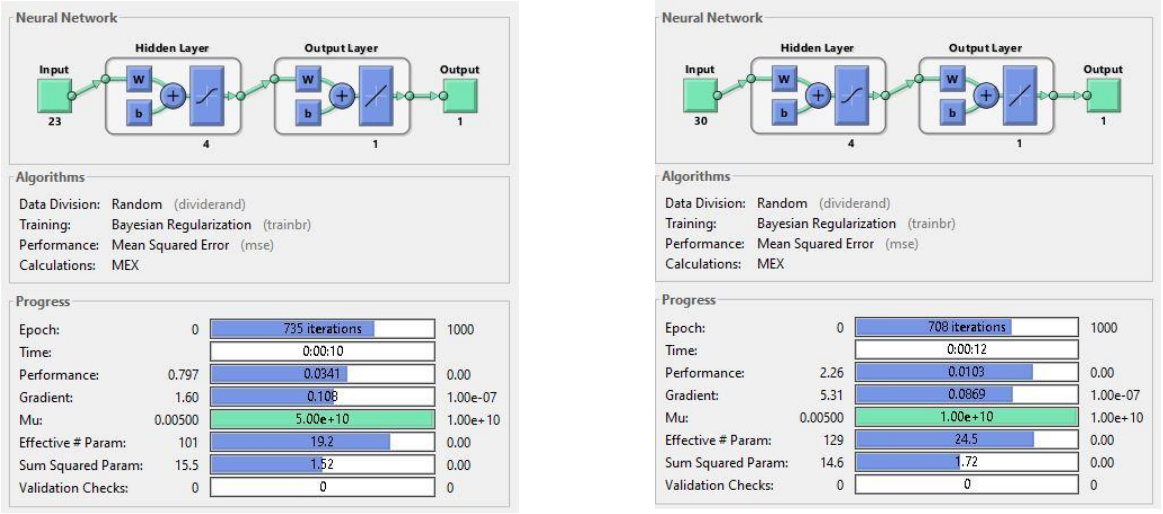
نمودار مدل ها به شرح زیر است:





نمودار۳۰: مدل شبکه عصبی رسم شده بر روی داده های نرمال شده با متود ۱ و متود ۴

روند انجام شبکه عصبی نیز به شکل زیر صورت گرفته است:



شکل۲: روند مدل کردن شبکه عصبی بر روی داده های نرمال شده با متود ۱ و متود ۴

با مقایسه مدل MLR و مدل های PCR و PLS و StepWise MLR در میابیم بر روی داده های این پروژه مدل شبکه عصبی بهتر نتیجه می دهد.

جدول ۶: مقایسه مدل های StepWise MLR و PCR و PLS و NN

ماتریس ورودی	Model	Components/ Features	Train		Test- MCCV		Test-LOO	
			R^2C	$RMSEC$	$R^2V - MCC$	$RMSEV - MCC$	$R^2V - LOO$	$RMSEV - LOO$
نرمال شده با متود ۱	NN	۳۰	۰,۸۶۲۲۱	۰,۵۰۷۴	۰,۷۳۸۰		۰,۷۳۸۰	
	StepWiseMLR	۹۲۹,۹۵۸,۲۹۴,۷۹۹,۸۴۱	۰,۷۸۱۵	۰,۲۳۸۴	۰,۶۸۳۷	۰,۳۰۳۹	۰,۶۸۲۵	۰,۲۹
	PCR	۱۴	۰,۶۵۲۲	۰,۳۰۵۳	۰,۱۹۶۶	۲,۰۳۱۲	۰,۲۵۹۶	۰,۷۰۱۲
	PLS	۷	۰,۹۴۱۲	۰,۲۴۸۵	۰,۱۴۳۰	۰,۵۸۴۸	۰,۱۲۱۱	۰,۵۳۸۱
نرمال شده با متود ۴	NN	۲۳	۰,۷۸۹۵	۰,۵۰۶۵	۰,۹۲۵۳		۰,۹۲۵۳	
	StepWiseMLR	۹۲۹,۹۵۸,۲۹۴,۷۹۹,۸۴۱	۰,۷۸۱۵	۰,۲۳۸۴	۰,۶۸۴۳	۰,۳۰۵۶	۰,۶۸۲۵	۰,۲۹
	PCR	۱۴	۰,۹۱۲۹	۰,۱۵۲۷	۰,۲۴۷۶	۱,۶۰۷۲	۰,۲۸۲۹	۰,۶۲۳۹
	PLS	۴	۰,۸۳۹۱	۰,۱۹۸۸	۰,۱۲۹۴	۰,۵۵۷۳	۰,۲۸۲۹	۰,۱۰۸۴

۳,۳,۵. Genetic Algorithm(GA):

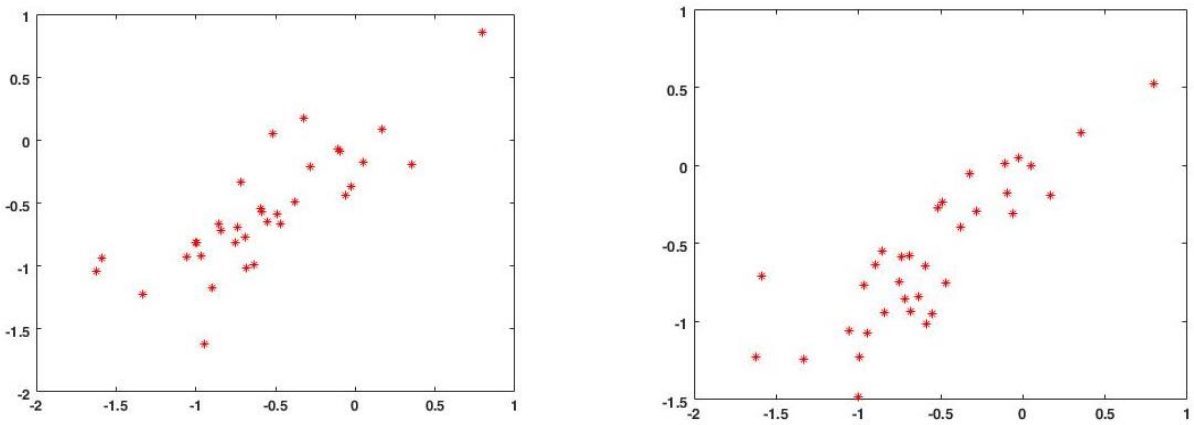
پس از اجرای مدل های PLS،PCR و NN حال این مدل ها را با الگوریتم ژنتیک ترکیب می کنیم. الگوریتم ژنتیک در هر با اجرا Population ای به صورت رندم ایجاد می کند و از بین آنها بهترین اعضا که objective function را کمینه می کنند انتخاب می کند و با crossover جمعیت جدیدی ایجاد و الگوریتم ادامه پیدا می کند. در هر سه الگوریتم GA_PLS و GA-MLR و GA-NN، تابع هدف $R^2 - 1$ است. کمینه کردن $R^2 - 1$ به معنای بیشینه کردن R^2 است.

در اجرای هر سه الگریتم ترکیبی تعداد ویژگی های انتخابی در هر مرحله برابر ۱۰ ویژگی و میزان جهش برابر ۰،۰۵ و تابع آن Uniform قرار داده شد. در اجرای GA-NN به دلیل کاهش هزینه زمان اجرایی، مقدار جمعیت در هر مرحله اجرا ۵۰ و تعداد مرحله اجرا برابر ۳۰۰ قرار داده شد. ورودی تمامی الگوریتم های ترکیبی داده های نرمال شده بدون کاهش ابعاد بوده است. نتایج زیر از اجرای الگوریتم های ترکیبی حاصل شد:

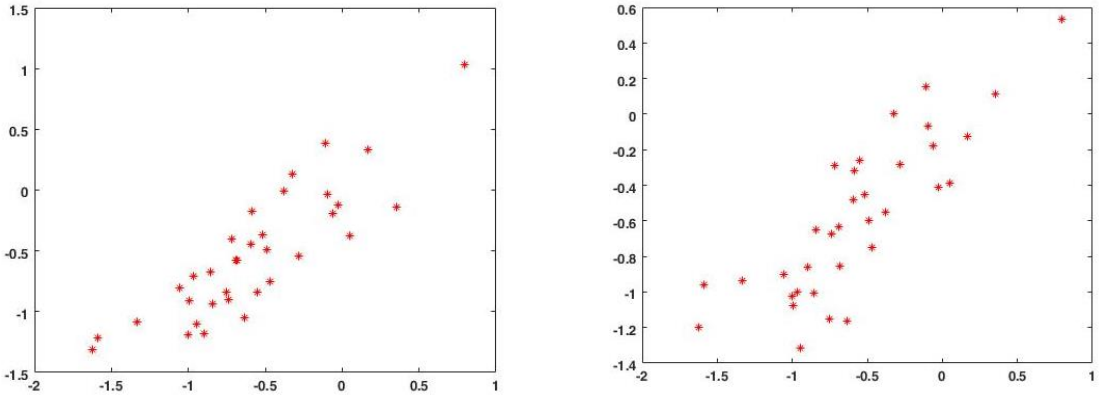
جدول ۷: مقایسه مدل های GA-PLSو GA-MLR و GA-NN

ماتریس ورودی	Model	Features	Components	Train		Test- MCCV		Test-LOO	
				R^2C	$RMSEC$	R^2V-MCC	$RMSEV-MCC$	R^2V-LOO	$RMSEV-LOO$
نرمال شده با متود ۱	GA-MLR	۶۸۰ , ۲۴۶ , ۶۵۷ , ۱۲۸۰ , ۷۹۲ ,۶۶۷ , ۱۱۶ , ۱۱۰۲ , ۱۰۱۸ , ۵۹۸	۱۰	۰،۸۴۶۱	۱۴e+۱،۱۲۶۸	۰،۶۳۵۰	۱۴e+۲،۱۷۴۱	۰،۷۱۶۴	۰،۲۷۳۹
	GA-PLS	۷۰۳ , ۷۴۳ , ۱۱۰۴ , ۸۰۸ , ۲۸۱ , ۳۳۰ , ۶۸۰ , ۱۵۹ , ۱۲۷۷ , ۷۱۷	۴	۰،۸۶۷۳	۰،۱۸۵۸	۰،۶۷۳۵	۰،۳۱۴۳	۰،۶۹۹۳	۰،۲۸۰۲
نرمال شده با متود ۴	GA-MLR	۹۳۰,۸۳۳,۷۴۹,۷۱۱, ۱۲۷۶,۱۰۹۷,۱۰۱۹, ۶۸۰,۱۲۳۲	۹	۰،۸۶۲۶	۸،۲۳۰۸e+۱۳	۰،۶۴۴۶	۶،۲۸۰۵□+۱۴	۰،۶۶۲۴	۰،۳۰۱۸
	GA-PLS	۹۰۷ , ۸۱۰ , ۸۶۵ , ۳۹۸ , ۱۰۱۹ , ۵۳۵ , ۷۱۹ , ۳۵۱ , ۷۱۶ , ۷۷۲	۴	۰،۸۷۵۶	۰،۱۷۹۹	۰،۶۶۹۱	۰،۳۱۸۸	۰،۷۴۶۴	۰،۲۷۲۴
داده اولیه	GA-NN	۱۲۴۲ , ۷۷۲ , ۱۸۵ , ۸۸۲ , ۳۱۳ , ۱۲۴۰ , ۲۰۴ , ۵۲۲ , ۱۱۴۵ , ۷۴۱	۱۰	۰،۹۰۸۹	۰،۱۷۴۲	۰،۸۹۴۷		۰،۸۹۴۷	

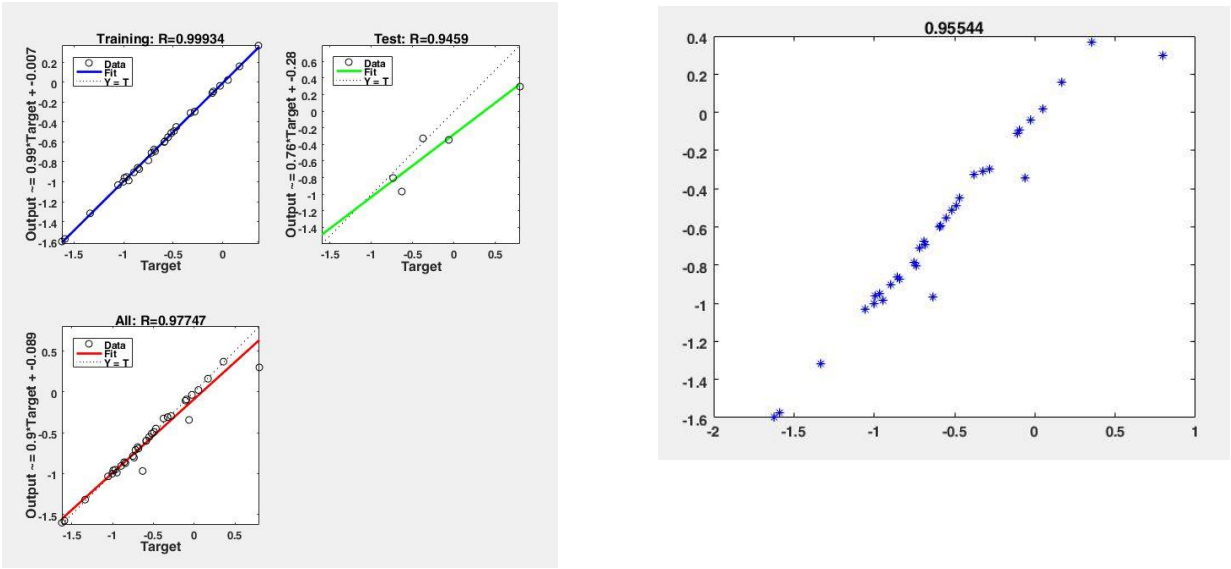
نمودار مدل ها به شرح زیر است:



نمودار۳۱: مدل GA-MLR رسم شده بر روی داده های نرمال شده با متود ۱ و متود ۴

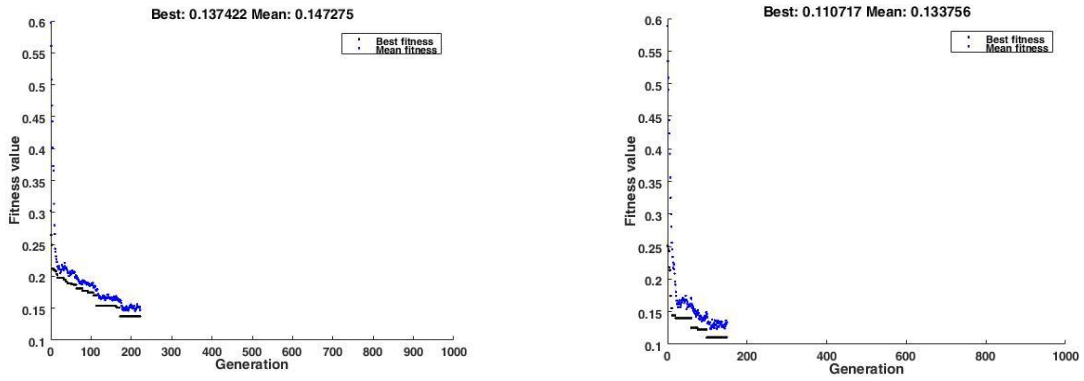


نمودار۳۲: مدل GA-PLS رسم شده بر روی داده های نرمال شده با متود ۱ و متود ۴

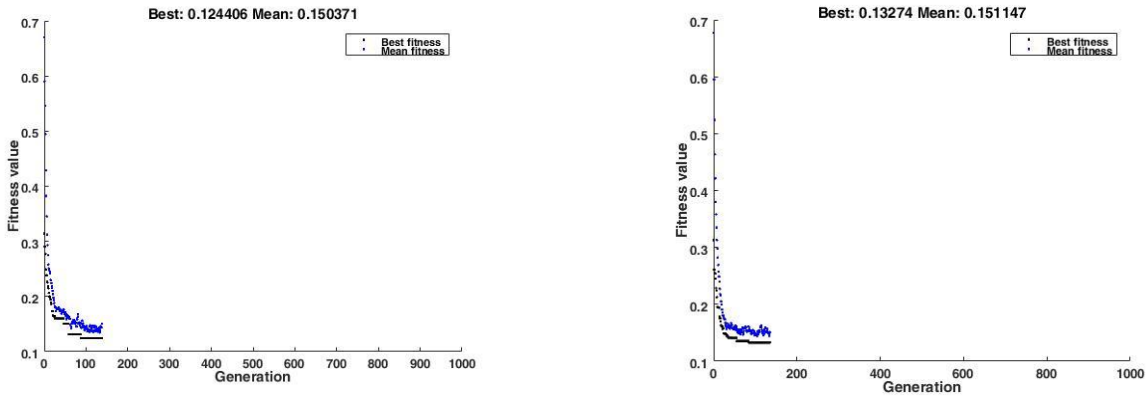


نمودار۳۳: مدل GA-NN رسم شده بر روی داده های اولیه به همراه مدل test و train

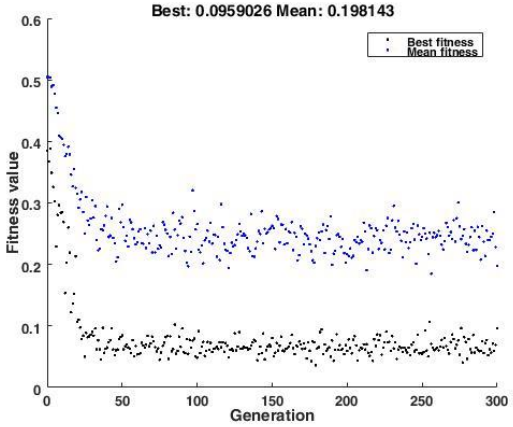
روند انجام الگوریتم ژنتیک به شرح زیر بوده است:



نمودار۳۴: روند طی شدن مراحل الگوریتم ژتتیک در GA-MLR



نمودار۳۵: روند طی شدن مراحل الگوریتم ژنتیک در GA-PLS



نمودار۳۶: روند طی شدن مراحل الگوریتم ژنتیک در GA-NN

حال به مقایسه مدل ها می پردازیم:

جدول ۸: مقایسه مدل های StepWise MLR و PCR و PLS و NN و GA-MLR و GA-PLS و GA-NN

ماتریس ورودی	Model	Components/ Features	Train		Test- MCCV		Test-LOO	
			R^2C	$RMSEC$	$R^2V - MCC$	$RMSEV - MCC$	$R^2V - LOO$	$RMSEV - LOO$
	GA-NN	۱۰	۰,۹۰۸۹	۰,۱۷۴۲	۰,۸۹۴۷		۰,۸۹۴۷	
نرمال شده با متود ۱	NN	۳۰	۰,۸۶۲۲۱	۰,۵۰۷۴	۰,۷۳۸۰		۰,۷۳۸۰	
	StepWiseMLR	۹۲۹,۹۵۸,۲۹۴,۷۹۹,۸۴۱	۰,۷۸۱۵	۰,۲۳۸۴	۰,۶۸۳۷	۰,۳۰۳۹	۰,۶۸۲۵	۰,۲۹
	GA-PLS	۴	۰,۸۶۷۳	۰,۱۸۵۸	۰,۶۷۳۵	۰,۳۱۴۳	۰,۶۹۹۳	۰,۲۸۰۲
	GA-MLR	۱۰	۰,۸۴۶۱	۱۴e+۱,۱۲۶۸	۰,۶۳۵۰	۱۴e+۲,۱۷۴۱	۰,۷۱۶۴	۰,۲۷۳۹
	PCR	۱۴	۰,۶۵۲۲	۰,۳۰۵۳	۰,۱۹۶۶	۲,۰۳۱۲	۰,۲۵۹۶	۰,۷۰۱۲
	PLS	۷	۰,۹۴۱۲	۰,۲۴۸۵	۰,۱۴۳۰	۰,۵۸۴۸	۰,۱۲۱۱	۰,۵۳۸۱
نرمال شده با متود ۴	GA-NN	۱۰	۰,۹۰۸۹	۰,۱۷۴۲	۰,۸۹۴۷		۰,۸۹۴۷	
	NN	۲۳	۰,۷۸۹۵	۰,۵۰۶۵	۰,۹۲۵۳		۰,۹۲۵۳	
	StepWiseMLR	۹۲۹,۹۵۸,۲۹۴,۷۹۹,۸۴۱	۰,۷۸۱۵	۰,۲۳۸۴	۰,۶۸۴۳	۰,۳۰۵۶	۰,۶۸۲۵	۰,۲۹
	GA-PLS	۴	۰,۸۷۵۶	۰,۱۷۹۹	۰,۶۶۹۱	۰,۳۱۸۸	۰,۷۴۶۴	۰,۲۷۲۴
	GA-MLR	۹	۰,۸۶۲۶	۸,۲۳۰۸e+۱۳	۰,۶۴۴۶	۶,۲۸۰۵□+۱۴	۰,۶۶۲۴	۰,۳۰۱۸
	PCR	۱۴	۰,۹۱۲۹	۰,۱۵۲۷	۰,۲۴۷۶	۱,۶۰۷۲	۰,۲۸۲۹	۰,۶۲۳۹
	PLS	۴	۰,۸۳۹۱	۰,۱۹۸۸	۰,۱۲۹۴	۰,۵۵۷۳	۰,۲۸۲۹	۰,۱۰۸۴

لازم به ذکر است با توجه به دقت بالای ارزیابی به روش مونت کارلو، ترتیب جدول ۸ بر اساس R^2 حاصل از ارزیابی مونت کارلو است.

References

[۱] Yunlei Hou, Liangyu Zhu, Zhiwei Li, Qi Shen, Qiaoling Xu, Wei Li, Yajing Liu, Ping Gong, "Design, synthesis and biological evaluation of novel bouchardatine analogs as potential inhibitors of adipogenesis/lipogenesis in ۳T۳-L۱ adipocytes," *European Journal of Medicinal Chemistry,* ۲۰۱۸.