

به نام خدا

## گزارش عملکرد مسابقه ی داده کاوی فناوری

اعضا :

مریم ابراهیم زاده

منیره صفری

سپیده ملانوروزی

یاسمن میرمحمد

فاطمه سادات هاشمی گلپایگانی

خرداد ۱۳۹۷

۱. مقدمه .....	۳
۲. خلاصه ی مختصری از روند اجرای کار .....	۴
۳. بخش اول : پردازش داده ها .....	۵
۴. بخش دوم : به دست آوردن مدل اولیه baseline .....	۲۳
۵. بخش سوم : پیاده سازی مدل svm با کرنل rbf .....	۲۹
۶. بخش چهارم : مدل خطی .....	۳۶
۷. بخش پنجم : استفاده از چند گام زمانی قبل تر برای تخمین .....	۳۶
۸. بخش ششم : نتیجه گیری .....	۳۶

مقدمه :

پیش بینی قیمت طلا یکی از مسائل پیچیده می باشد و عامل های مختلفی روی آن تاثیر می گذارد. عواملی چون تورم، سیاست خارجی ، میزان تقاضا، درآمد حاصل از فروش نفت و .... . بعد از مطالعات انجام شده به این مسئله به یک مسئله سری زمانی مدل شد و مدلی هایی روی آن امتحان شد. با تصویر سازی داده های مربوط به طلا به نظر می آمد که قیمت آن در هر مقطع زمانی تابعی از زمان نبوده و به صورت شهودی تابعی از مقدار قیمت در مقطع زمانی قبلی است. از این جهت قیمت در هر لحظه را تابعی از قیمت در زمان قبلی در نظر گرفته شد. با این روش این مسئله به یک مسئله supervised تبدیل شد. بخشی از داده ها به عنوان داده تست جدا شدند و مدل بر اساس بقیه ی داده ها آموزش داده شد. خطای هر مدل با مقایسه مقدار تخمین زده شده توسط مدل آموزش داده شده برای داده های تست و مقدار واقعی آن ها اندازه گیری شد.

خلاصه ی مختصری از روند کار :

ابتدا فایل صورت سوال خوانده شد و پس از بررسی کلی اطلاعات داده شده و صورت مسئله اعضا به مطالعه ی مدل های مناسب برای این نوع مسائل پرداختند. مدل های مطالعه شده شامل مدل شبکه عصبی lstm، مدل svm، مدل anfis، مدل normal و مدل خطی بود که تصمیم بر این شد ابتدا از مدل های ساده شروع کرده و در هر مرحله مدل قبلی بهبود داده شود و یا از مدل پیچیده تری استفاده شود . برای زبان برنامه نویسی نیز با توجه به مهارت اعضا و کتابخانه ها امکانات مورد نیاز زبان پایتون انتخاب شده است .

مراحل بعدی کار را می توان به بخش های زیر تقسیم کرد:

۱. بررسی فایل داده ها
۲. تبدیل داده ها به داده های نظارت شده ( supervised )
۳. به دست آوردن مدل اولیه baseline و محاسبه ی خطای آن
۴. اجرای مدل svm و محاسبه ی خطا و تصویرسازی
۶. اجرای مدل خطی بر روی کل داده ها
۷. اجرای مدل خطی بر روی داده های بازه بندی شده
۸. مقایسه مدل های مختلف روی موضوعات مختلف و انتخاب بهترین مدل برای هر ارز یا فلز گرانبها

## بخش اول :

پردازش داده ها :

ابتدا داده های فایل ticker بررسی شده اند .

تمامی ویژگی های داده شده در این فایل (ستون ها) از a0 تا a8 نام گذاری شده اند.

a0: زمان ثبت اطلاعات که با رزولوشن ۵ دقیقه جمع آوری شده است

a1: قیمت حداکثر سفارش ثبت شده برای فروش

a2: حجم حداکثر سفارش ثبت شده برای فروش .

a3: قیمت حداکثر سفارش ثبت شده برای خرید

a4: حجم حداکثر سفارش ثبت شده برای خرید .

a5: قیمت آخرین معامله صورت گرفته در بازار

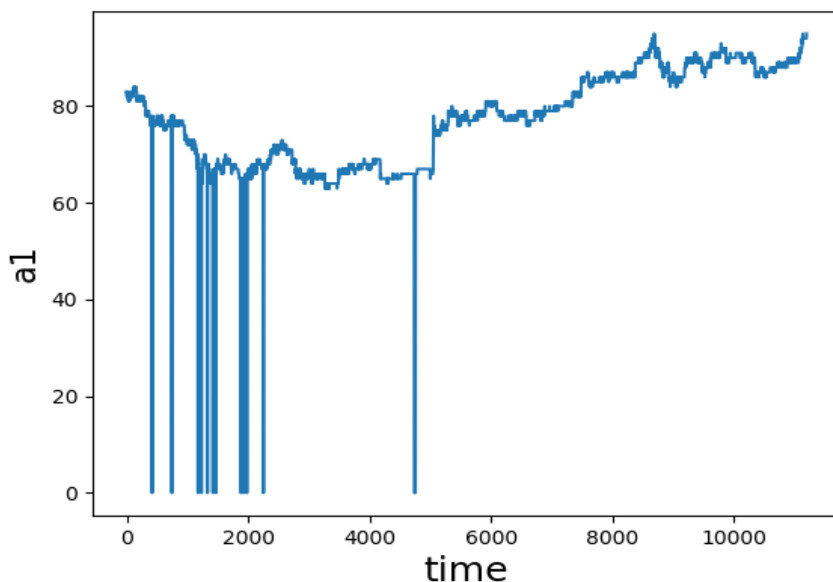
a6: حجم معاملات روزانه .

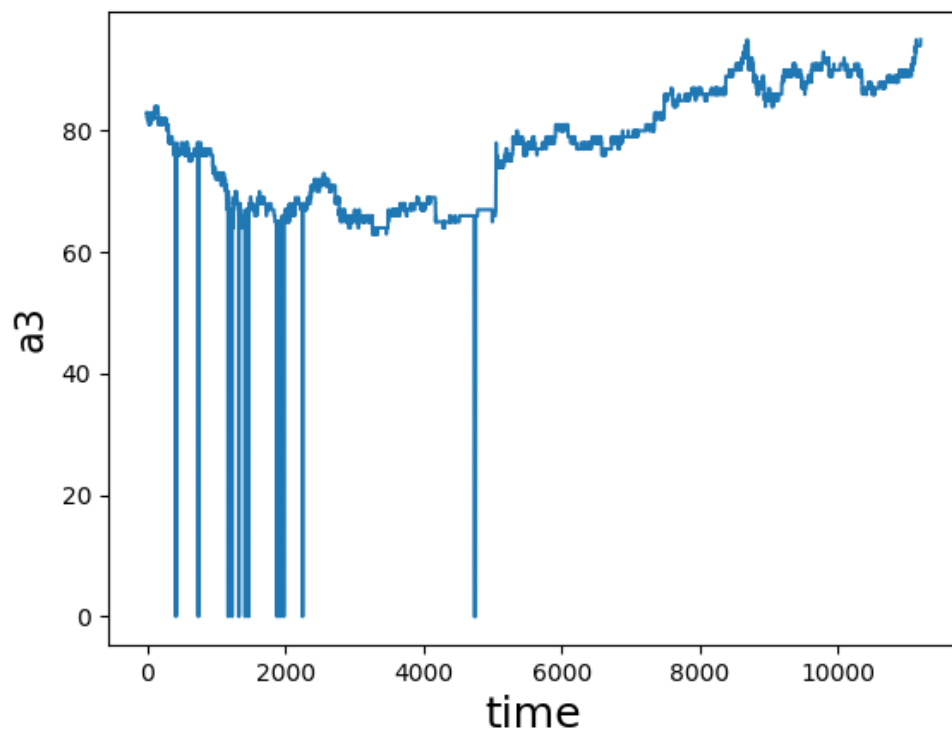
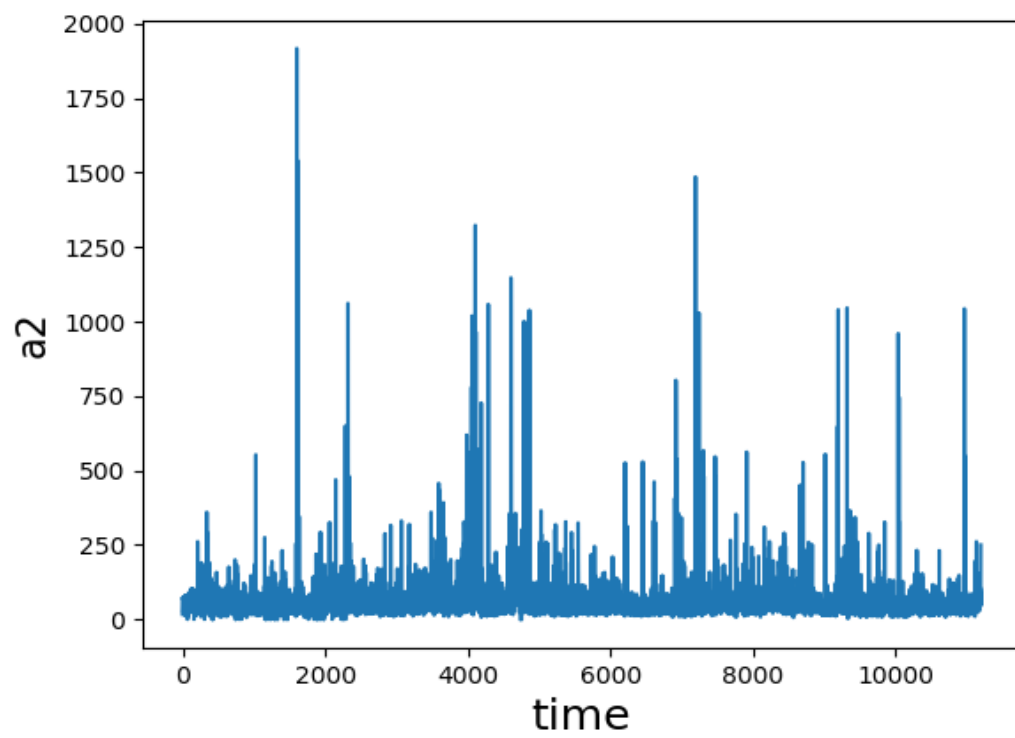
a7: حداکثر قیمت روزانه .

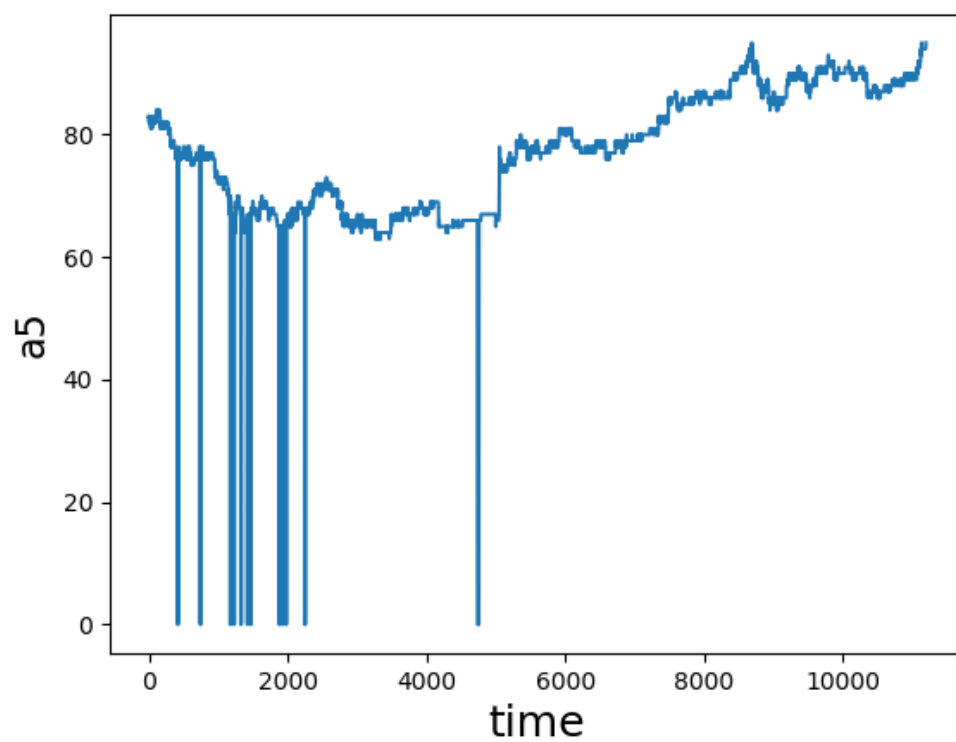
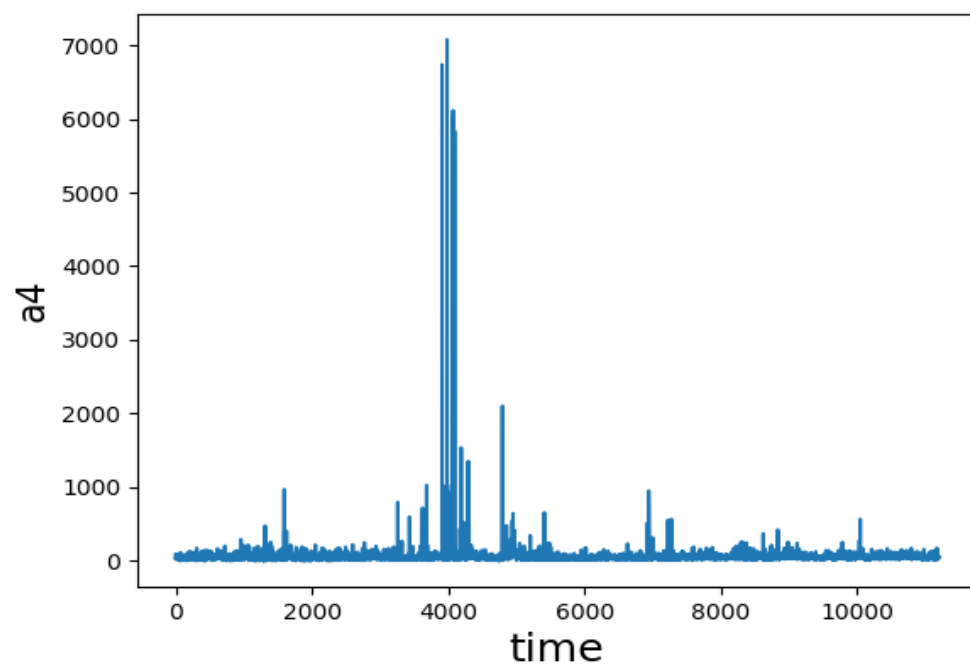
a8: حداقل قیمت روزانه

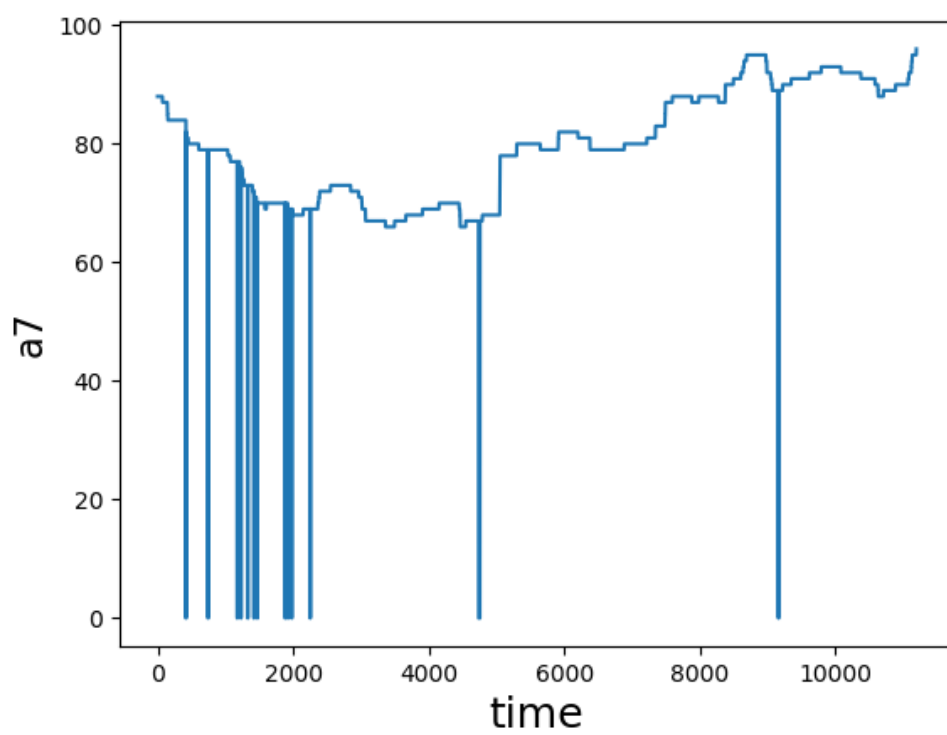
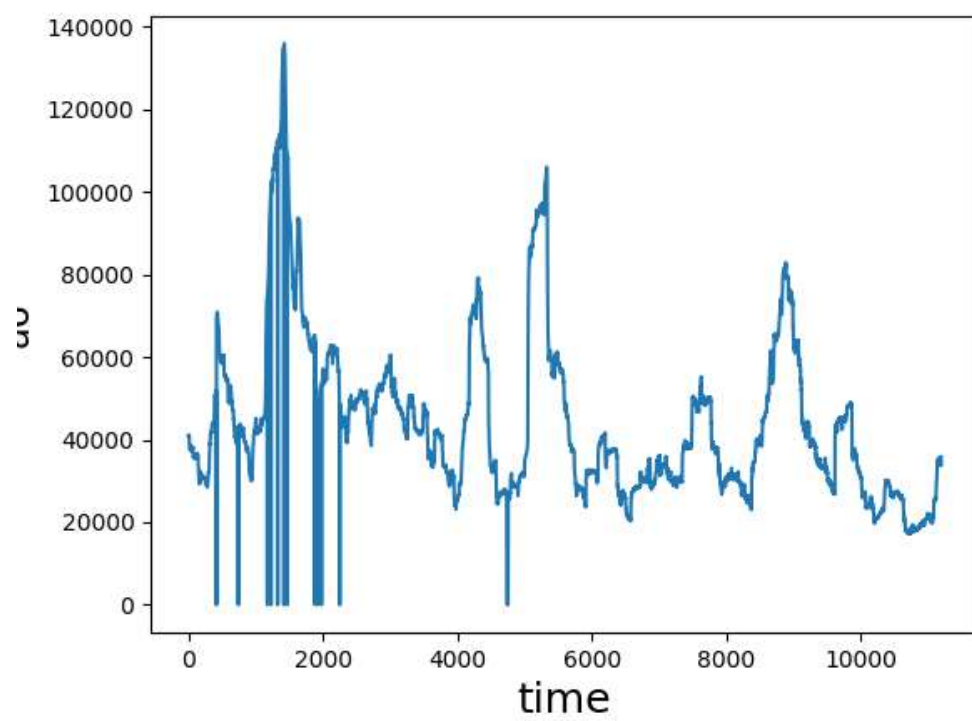
همه داده ها بر اساس زمان(a0) کشیده شده اند .

(منظور از b1 هم همین ویژگی a1 است ولی برای جنس b و به همین ترتیب )

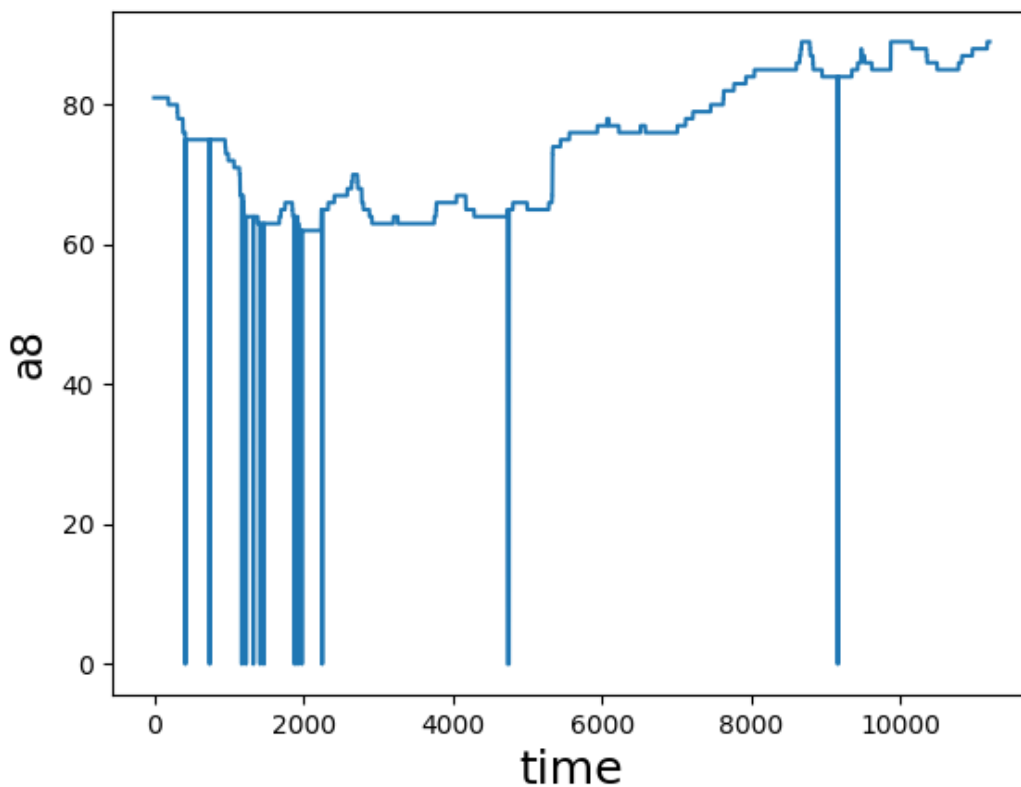






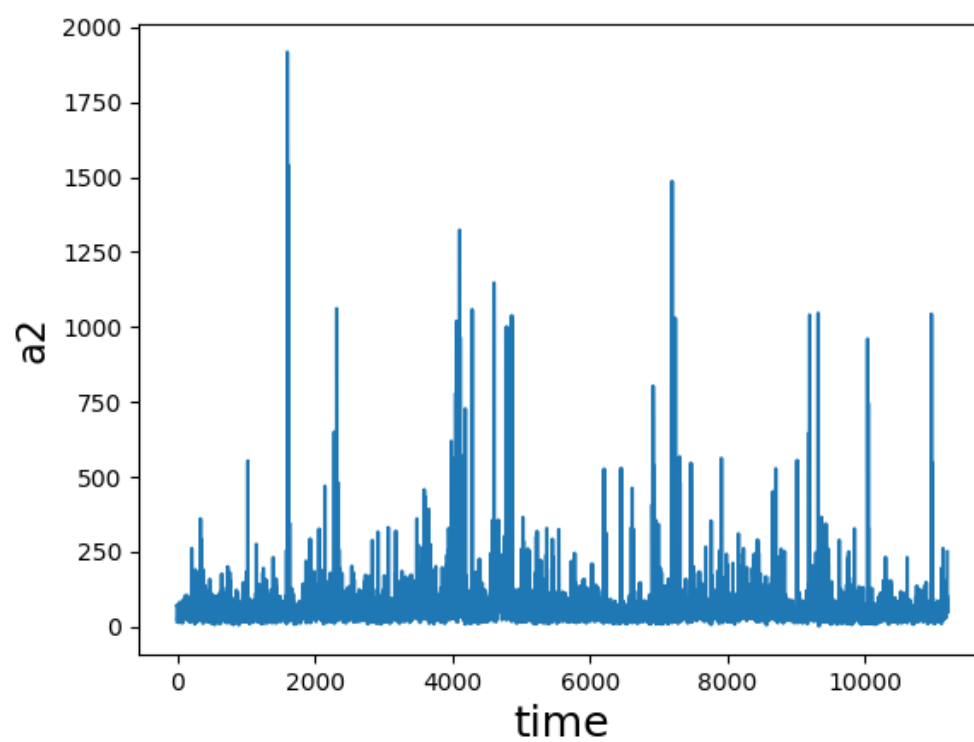
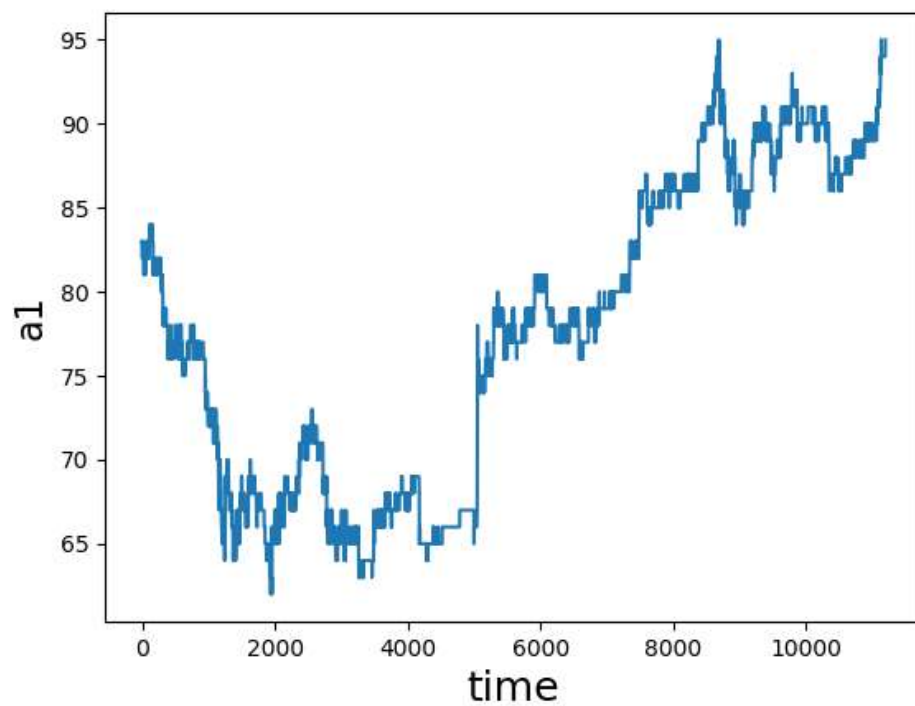


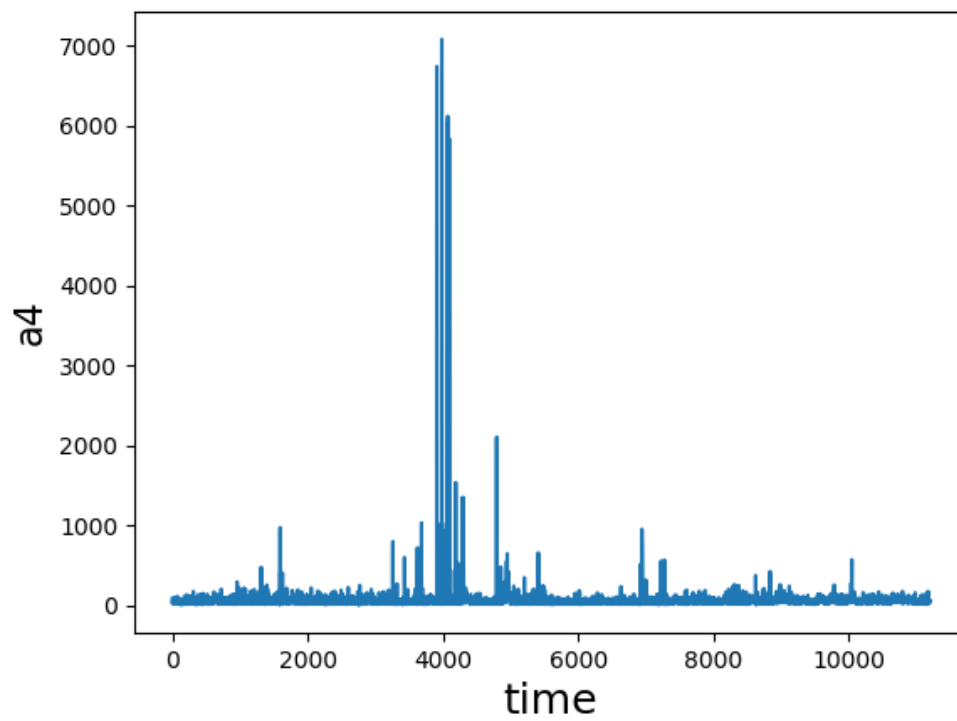
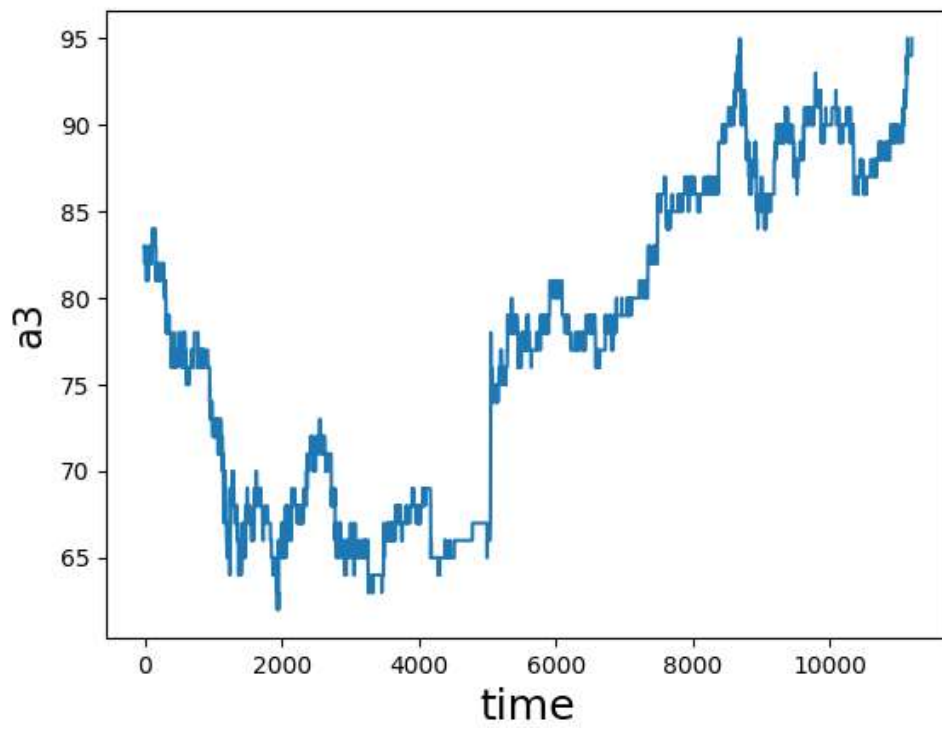


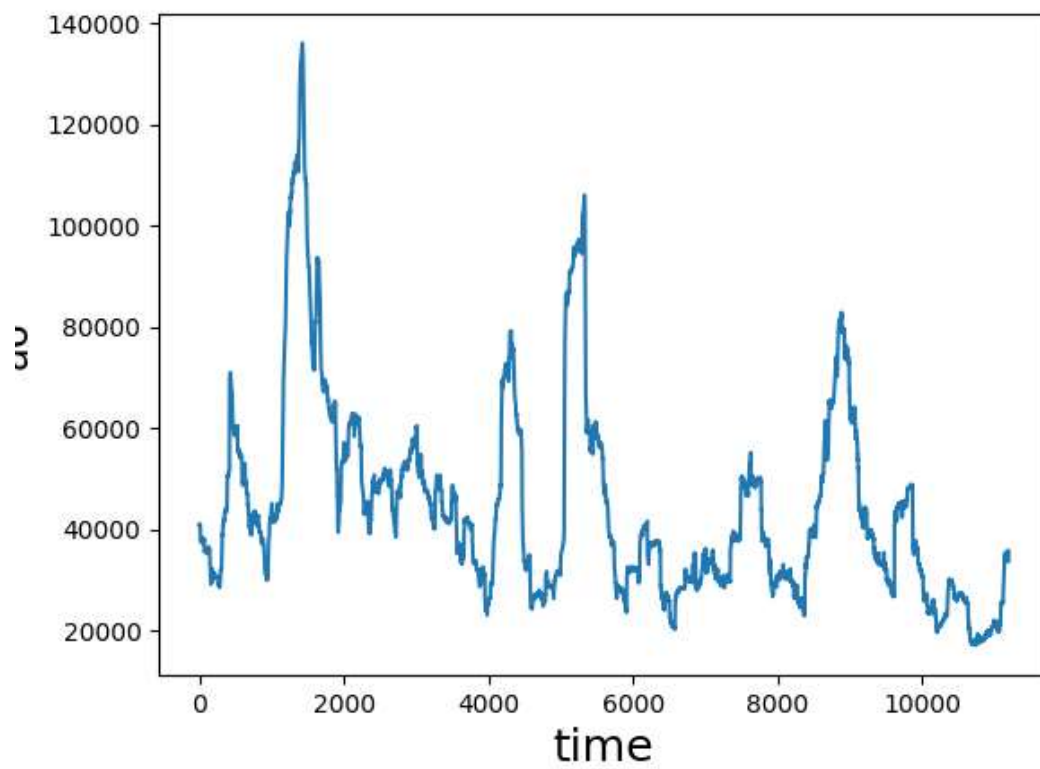
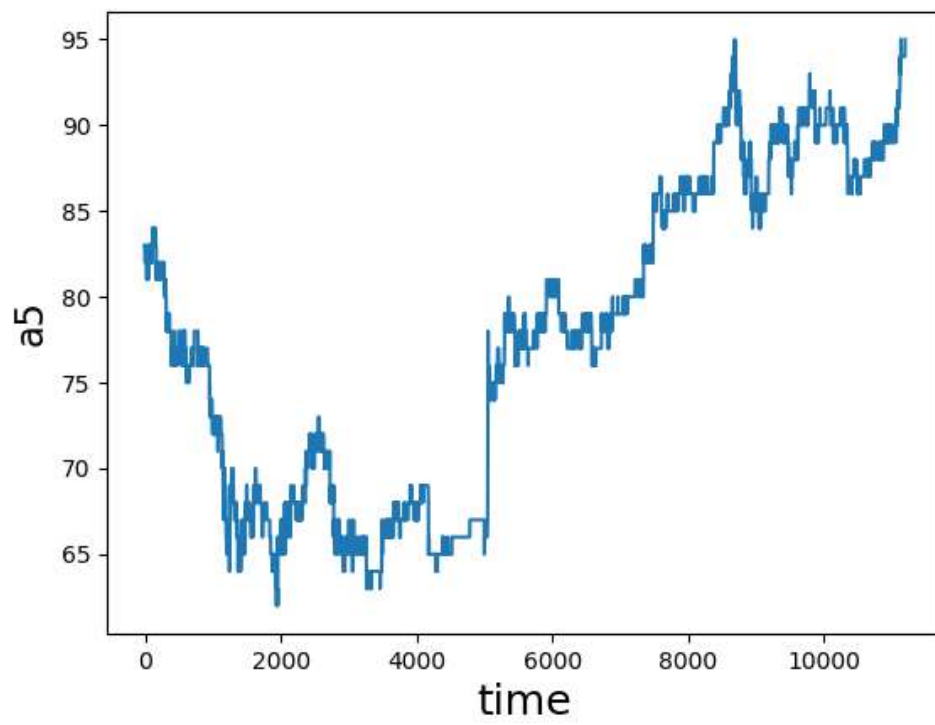


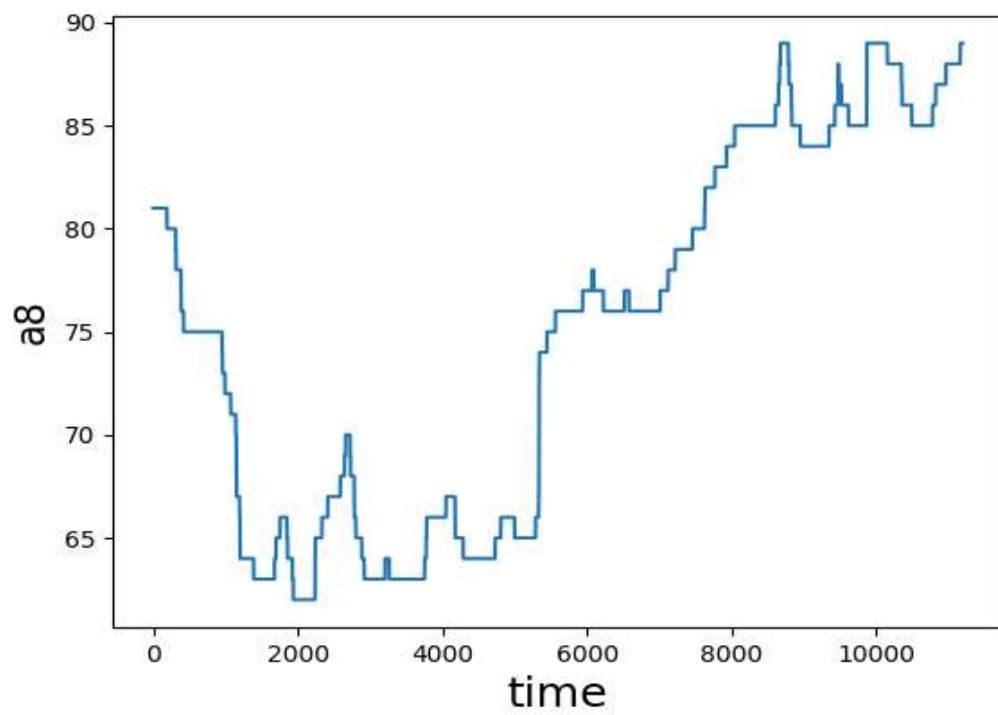
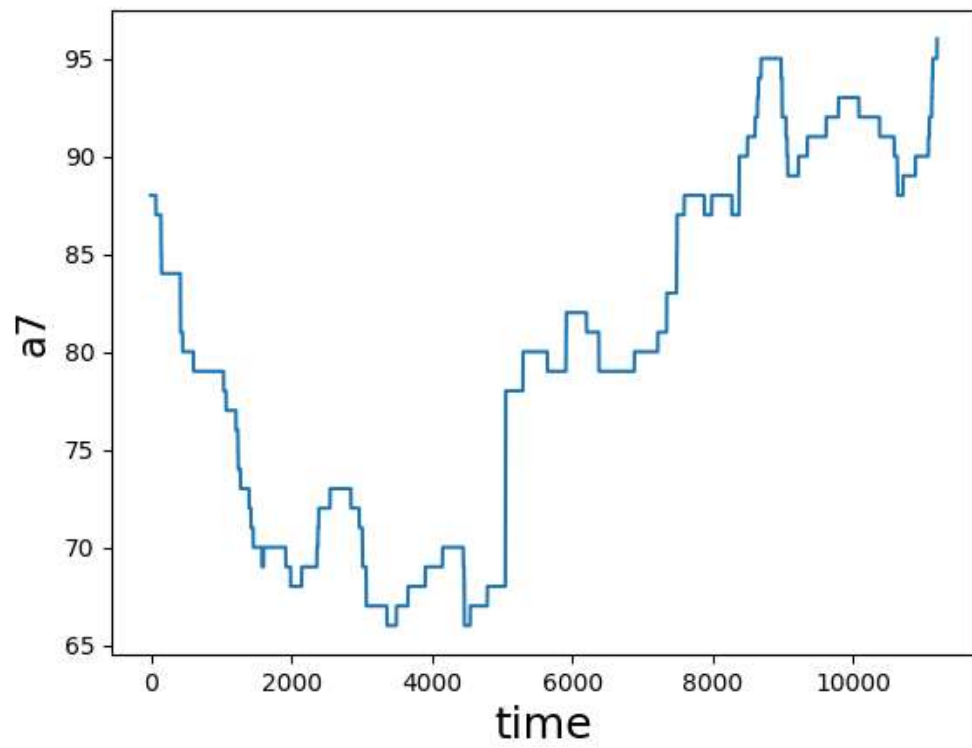
همانطور که از شکل های کشیده شده نتیجه گیری می شود نمودار تغییرات  $1a$  و  $3a$  و  $5a$  و  $7a$  و  $8a$  بر حسب زمان مشابه هم هستند.

همچنین از شکل های کشیده شده نتیجه گیری می شود که فایل داده ها دارای داده های پرتی یا داده های اشتباهی هستند که باعث ایجاد خطا در مراحل بعدی میشوند در نتیجه آن ها را از اطلاعات حذف کردیم و همانطور که در شکل های زیر مشاهده می شود تغییرات ناگهانی داده ها از بین رفت.



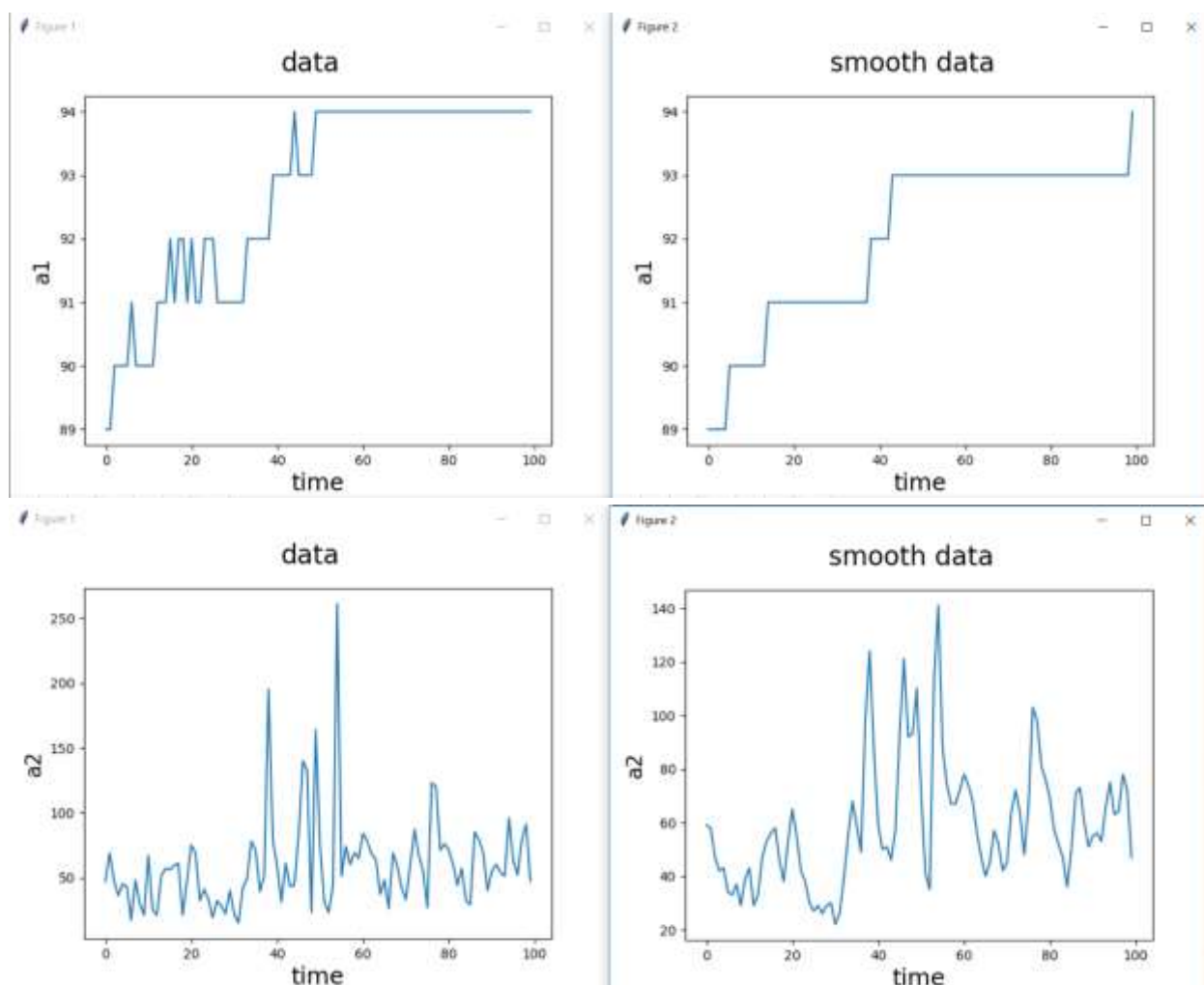


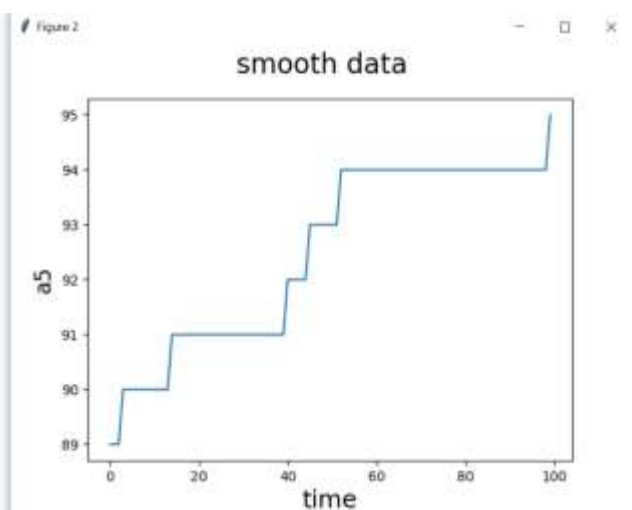
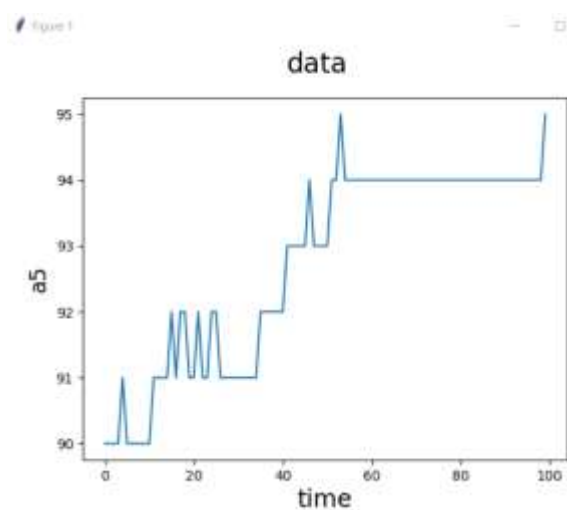
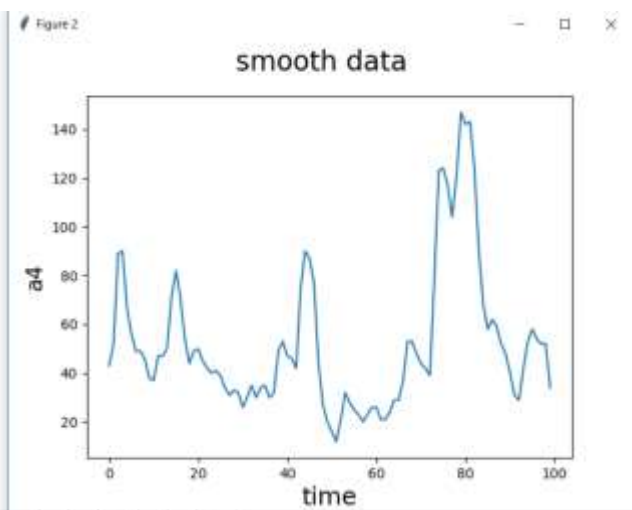
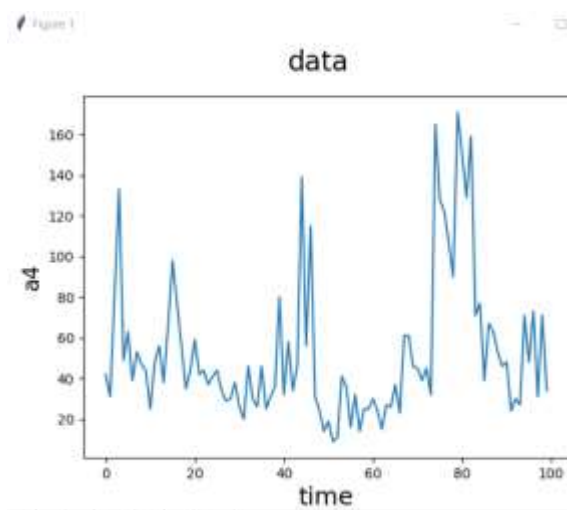
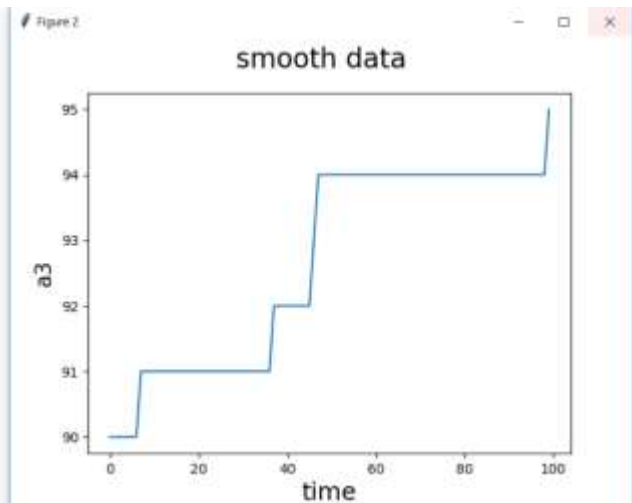
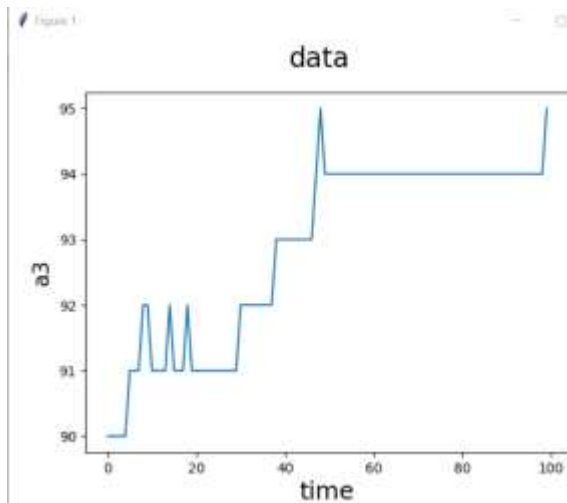


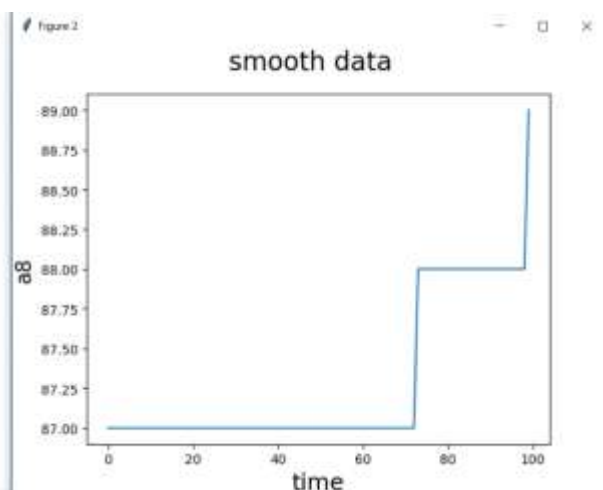
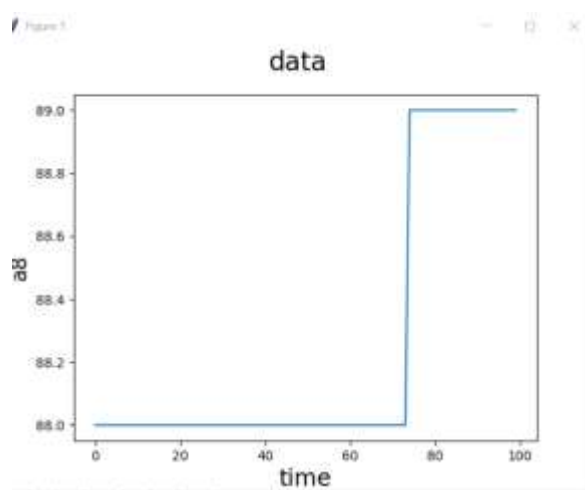
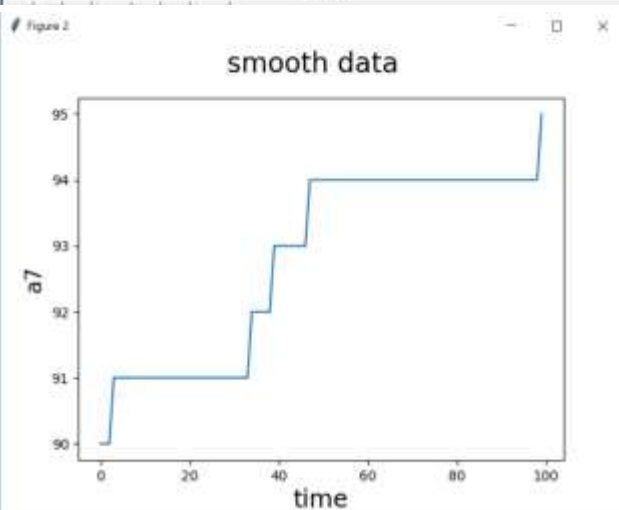
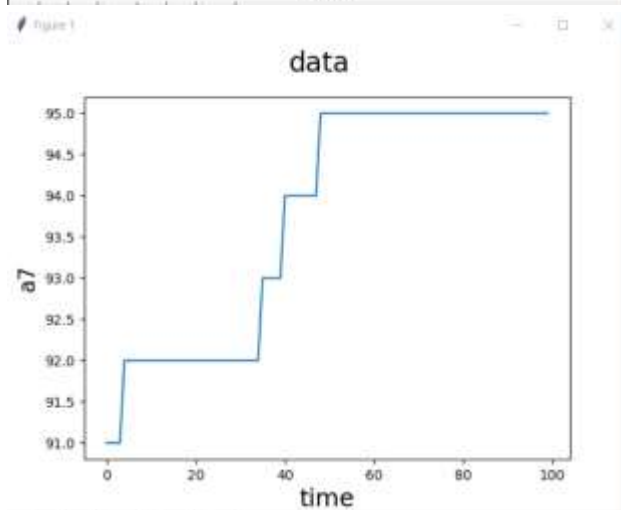
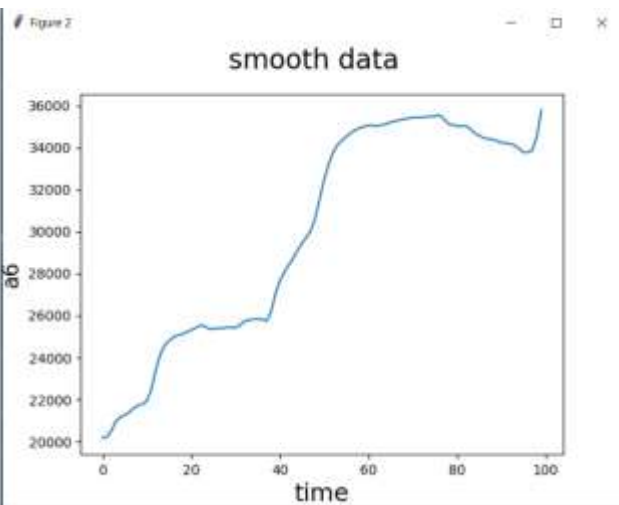
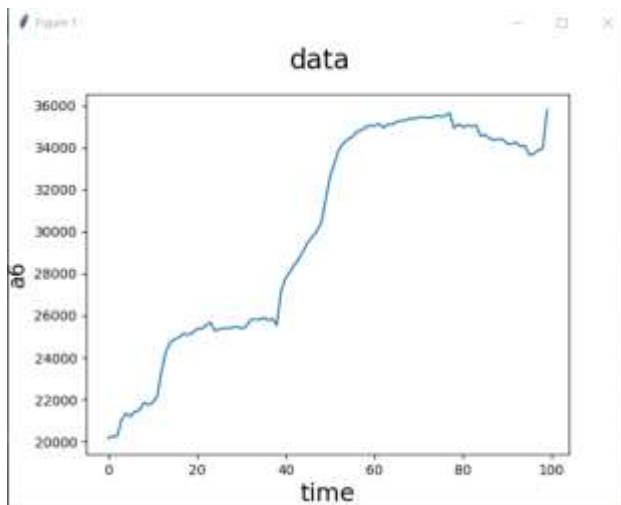


هموارسازی داده ها :

یکی دیگر از کار های انجام شده روی داده ها هموارسازی داده ها بود که به این صورت انجام شد که مقدار هر نقطه برابر میانگین مقدار آن نقطه و یک مقدار قبل آن و یک مقدار بعد آن قرار می دهید. شکل ۱۰۰ داده ی آخر پس از هموار سازی (برای بهتر نشان داده شدن تغییرات پس هموارسازی همه ی داده ها رسم نشده است) :





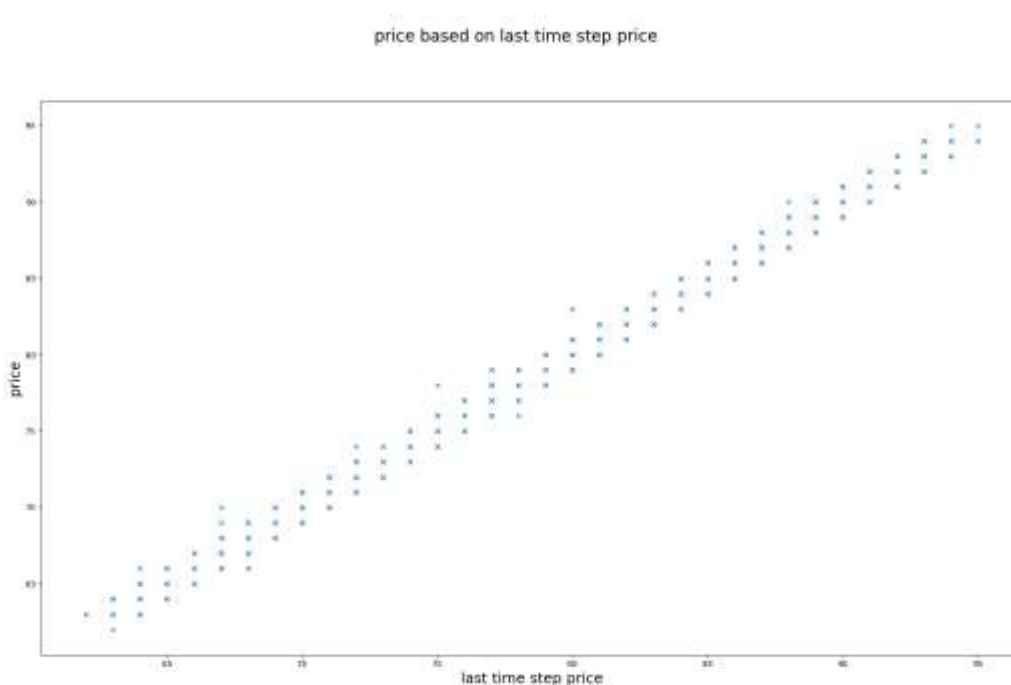




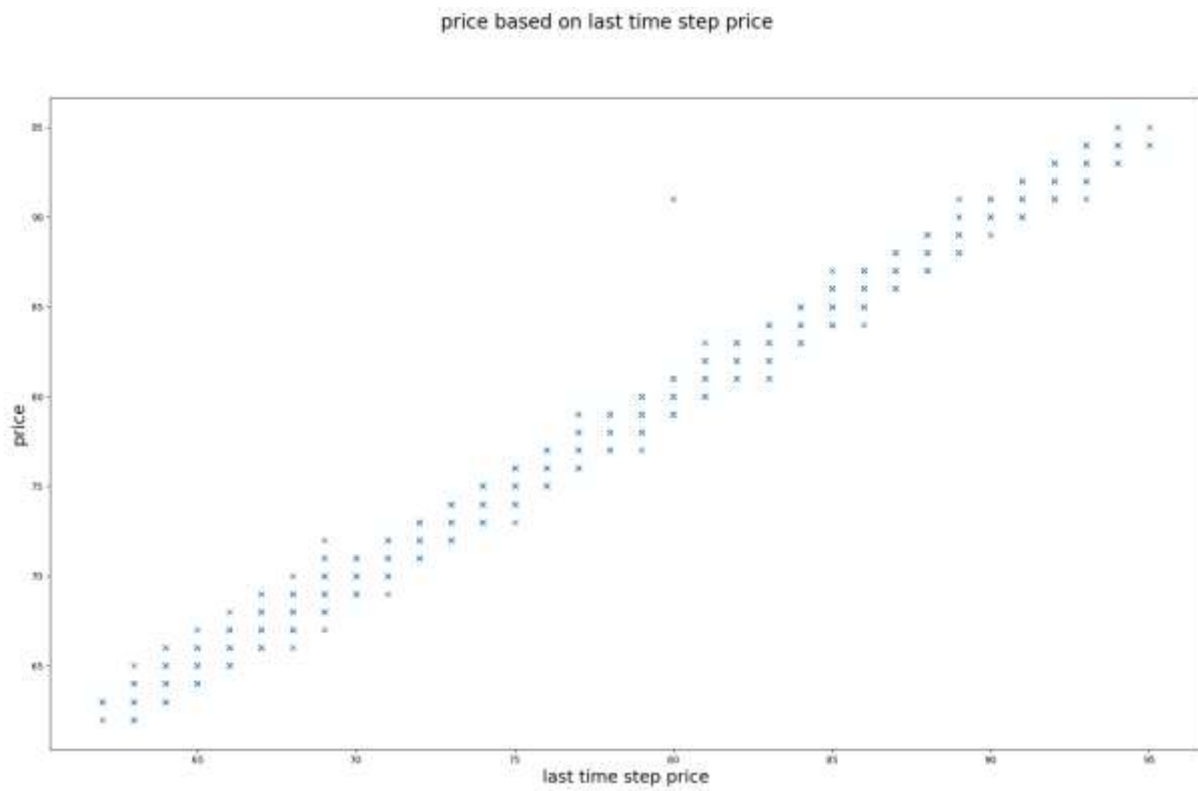
تصویر سازی داده ها:

همان طور که پیش تر گفته شد داده ها به این صورت مدل شدند که مقدار قیمت در هر لحظه تابع مقدار قیمت در لحظه قبلی در نظر گرفته شد. بر همین اساس داده ها رسم کردیم و به شکل های زیر رسیدیم:

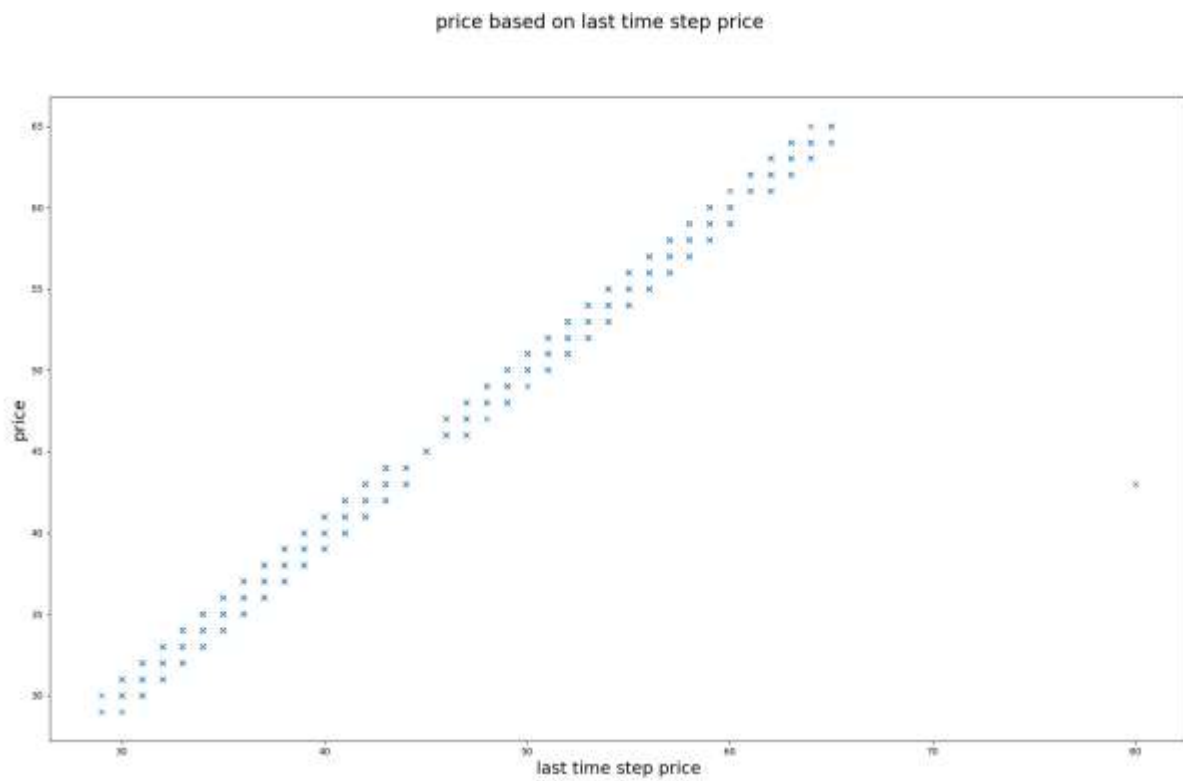
A\_ticke (a1):



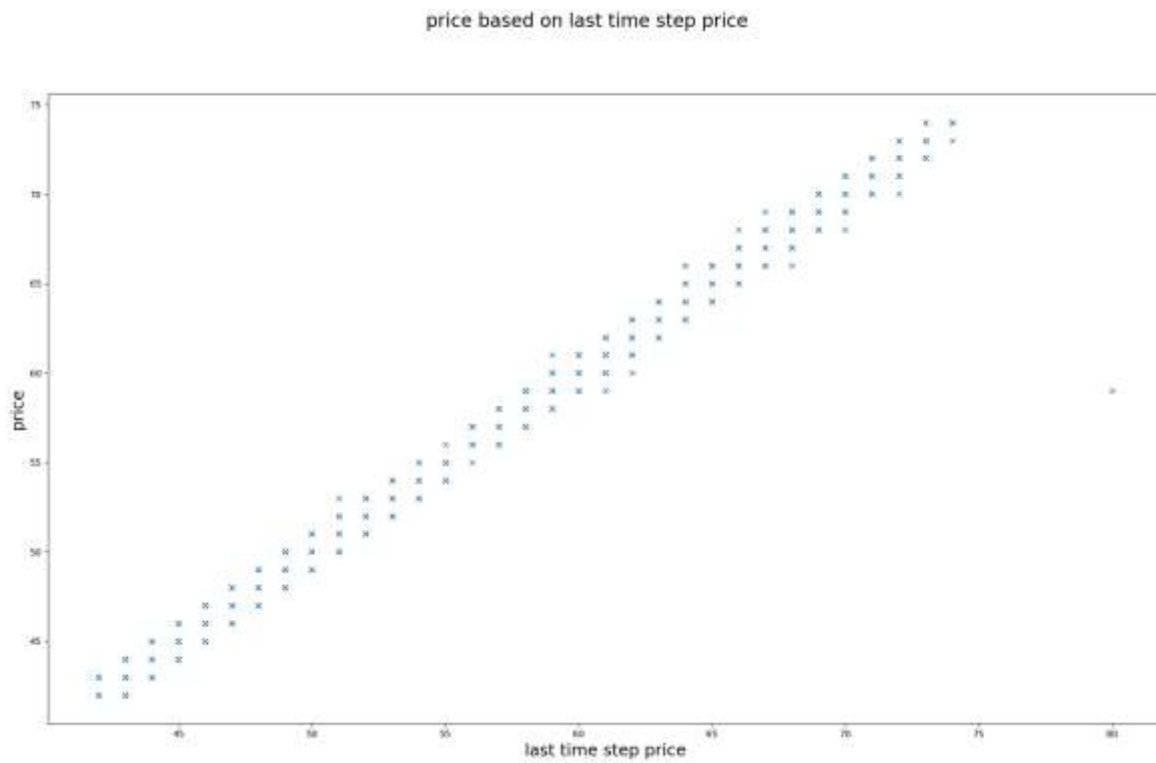
B\_ticker(b1):



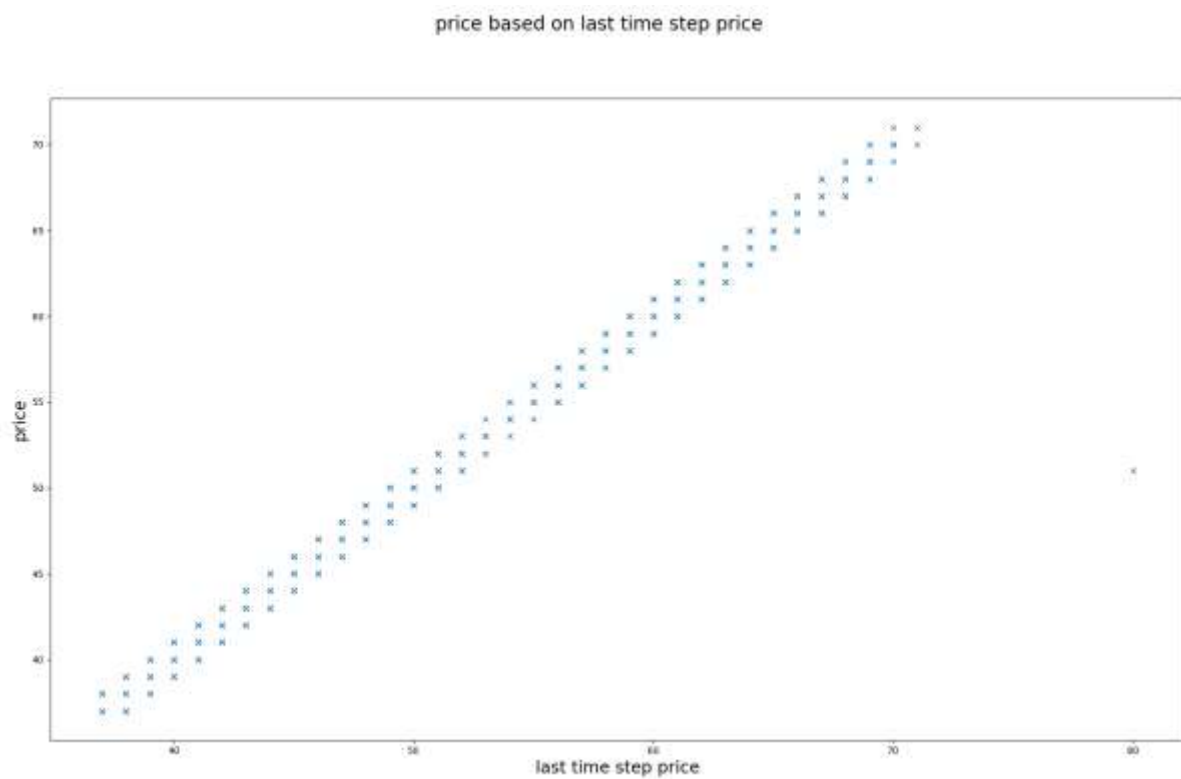
C\_ticker(c1):



D\_ticker(d1):

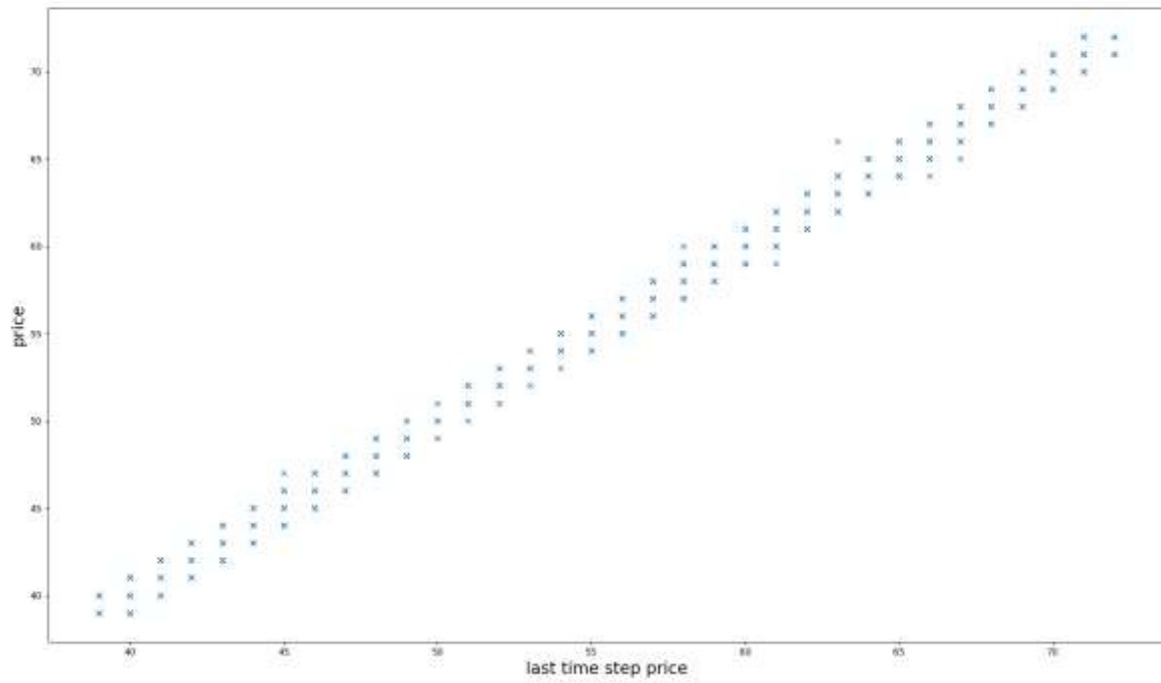


E\_ticker(e1):



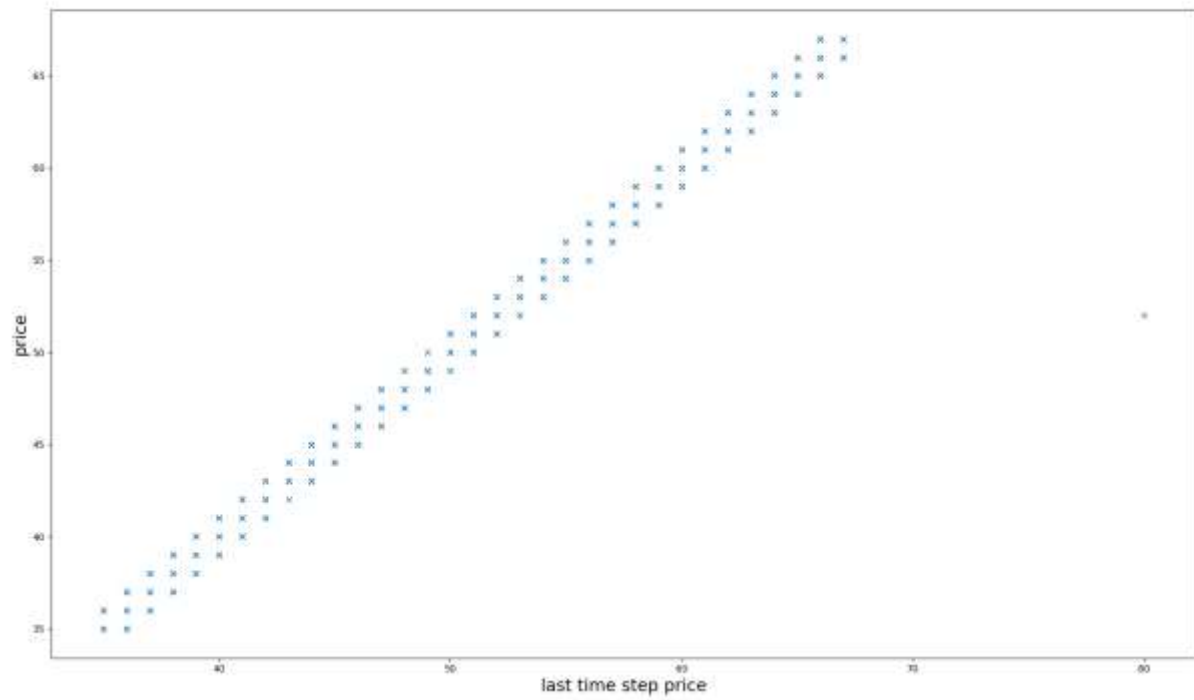
F\_ticker(f1):

price based on last time step price

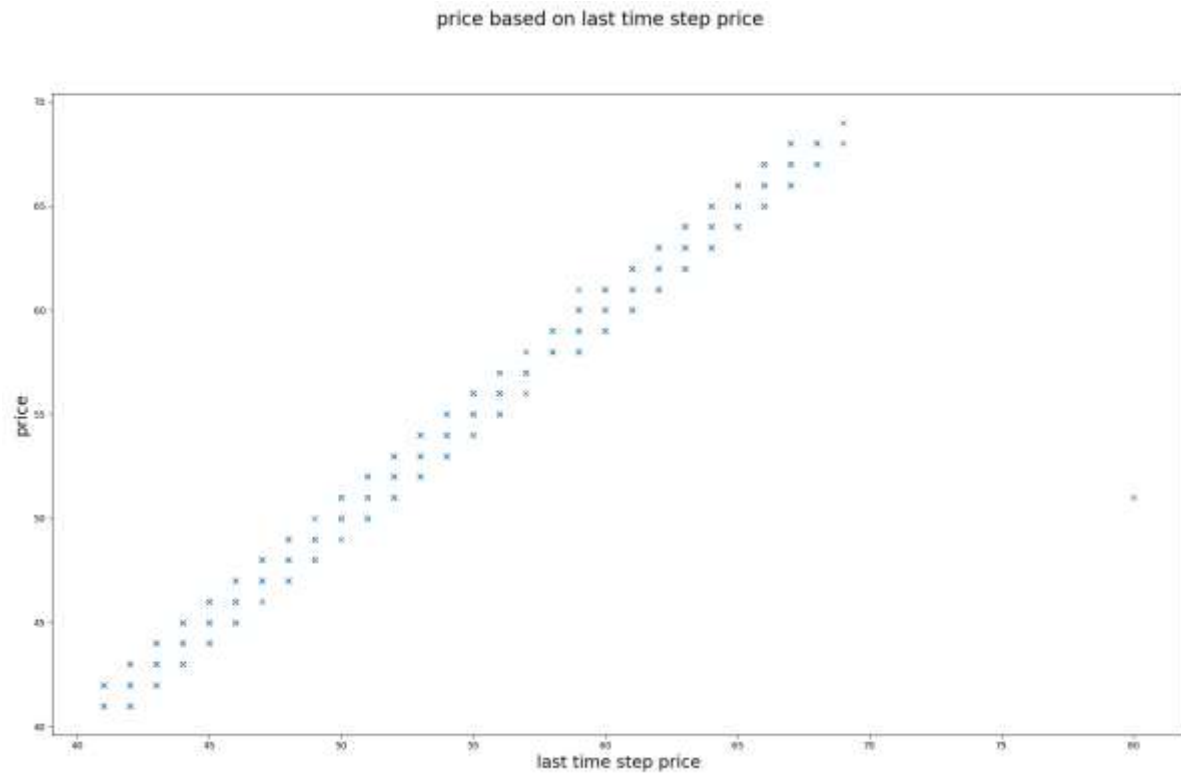


G\_ticker(g1):

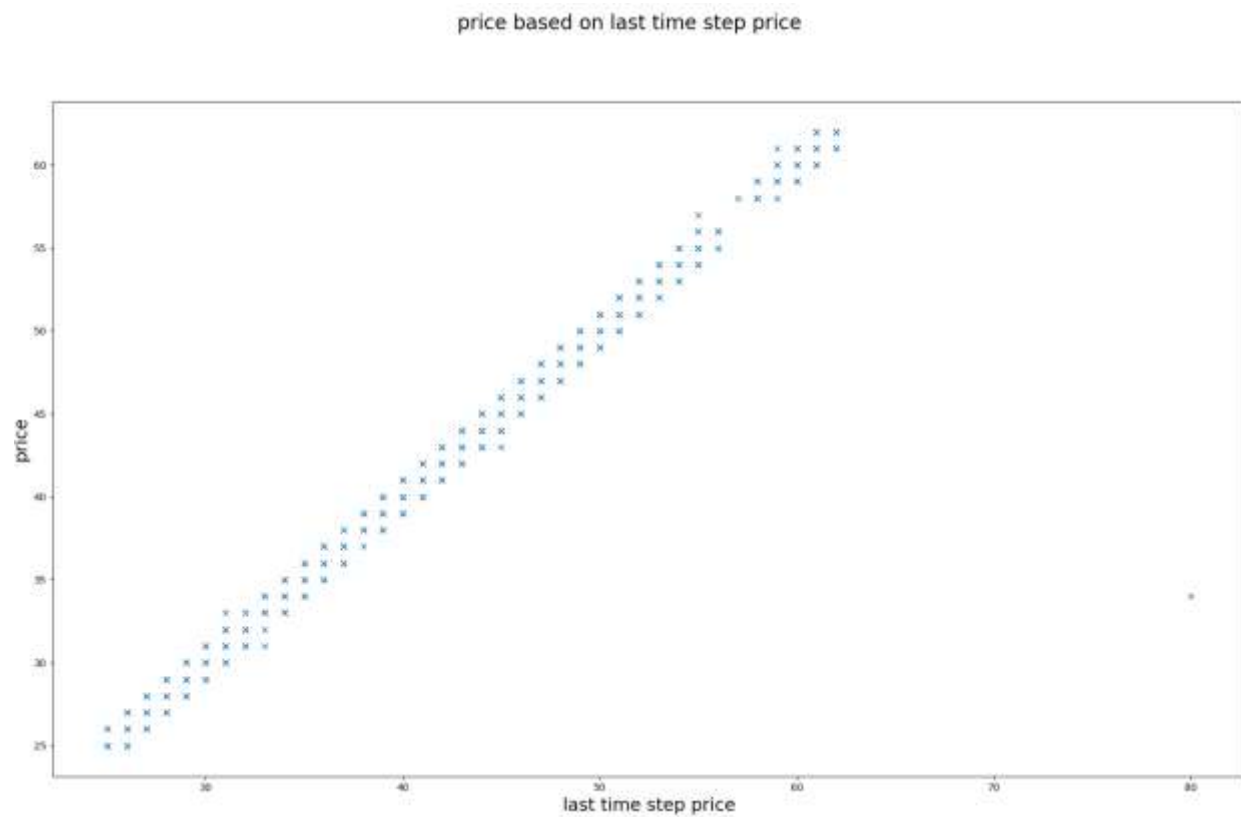
price based on last time step price



H\_ticker(h1):



I\_ticker(i1):



همان طور که مشاهده می شود داده ها حالت خطی دارند و برای همین پیش بینی می شود که مدل های خطی نتیجه خوبی داشته باشند.

## بخش دوم :

به دست آوردن مدل اولیه baseline و محاسبه ی خطای آن :

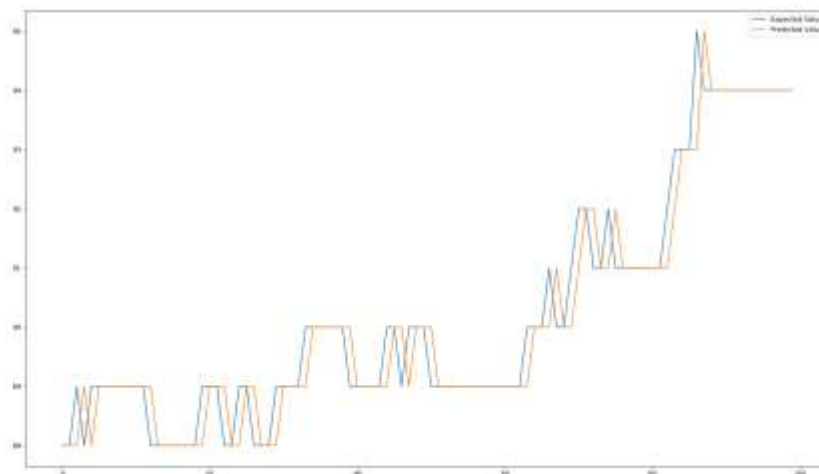
به طور کلی برای اجرای این الگوریتم ها و پیدا کردن مدلی برای تخمین قیمت آینده نیاز به ایجاد داده های train و test داشتیم که برای این کار بعد از خواندن اطلاعات فایل اول به صورت رندوم ۲۰ درصد از داده ها به عنوان test و ۸۰ درصد دیگر به عنوان train انتخاب شدند.

قبل از شروع تست کردن مدل های مختلف روی داده نیاز بود که یک baseline برای تخمین خود داشته باشیم. بدین منظور که بتوانیم دیگر مدل ها را با آن مقایسه کنیم و از مقایسه آن ها با این مدل میزان کارایی آن ها را مورد بررسی قرار دهیم .

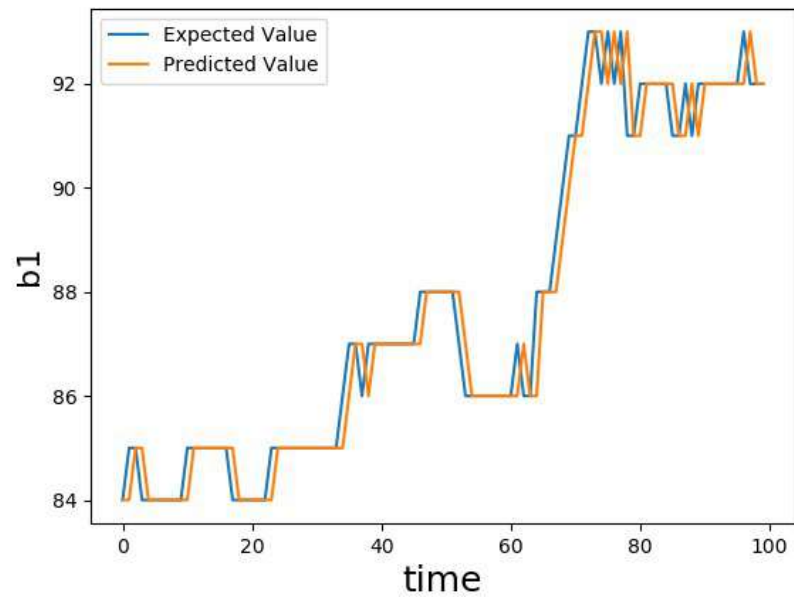
برای این منظور از یک مدل ساده که برای تخمین داده های سری زمانی استفاده می شود و مقدار قیمت در هر گام زمانی را برابر قیمت در گام قبلی در نظر می گیرد استفاده کرده ایم.

برای بررسی مدل تصویر داده های اصلی و مقدار تخمین زده شده کشیده شد و همچنین میزان خطای rmse مدل محاسبه شده است

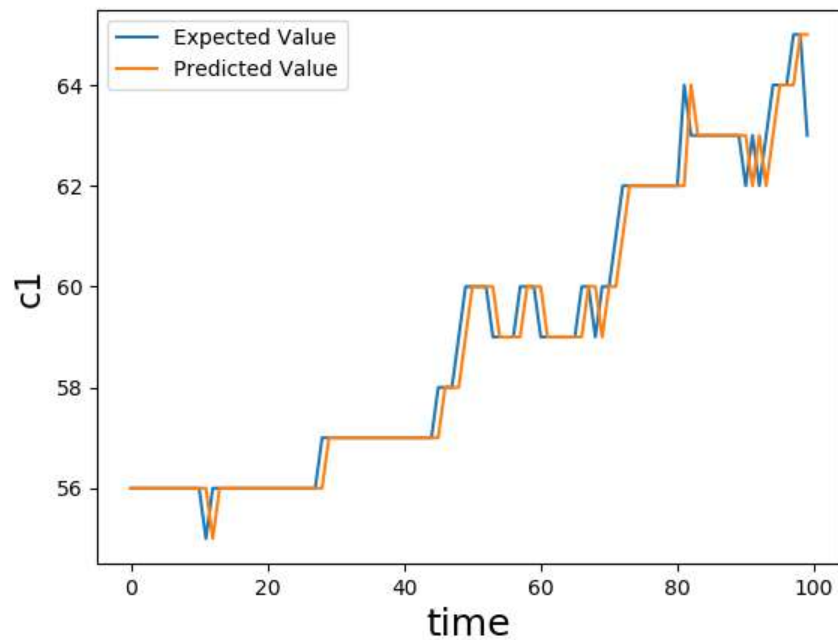
تصویر ۱۰۰ تا از داده های تست فایل A\_ticker برای baseline :



تصویر ۱۰۰ تا از داده های تست فایل B\_ticker برای baseline :

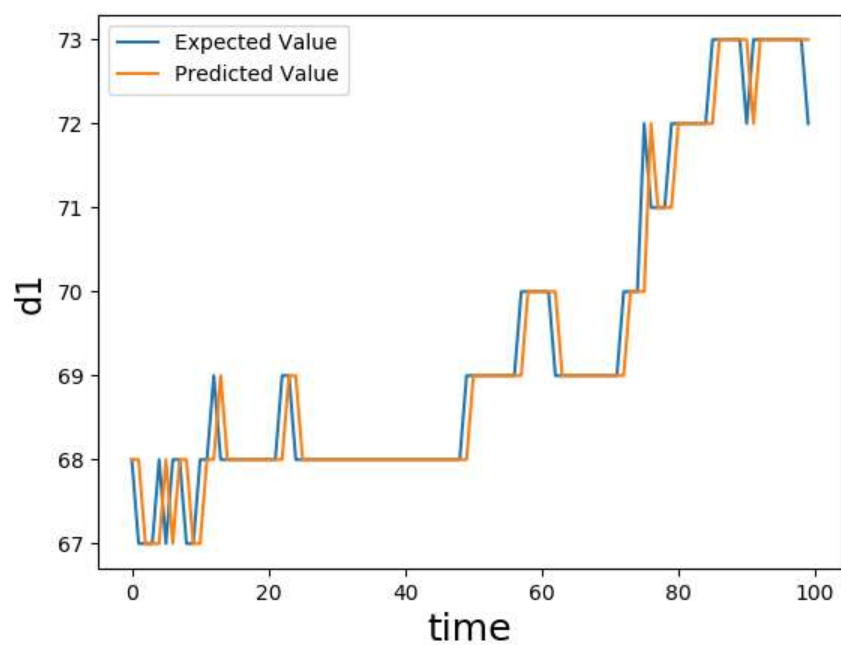


تصویر ۱۰۰ تا از داده های تست فایل C\_ticker برای baseline :

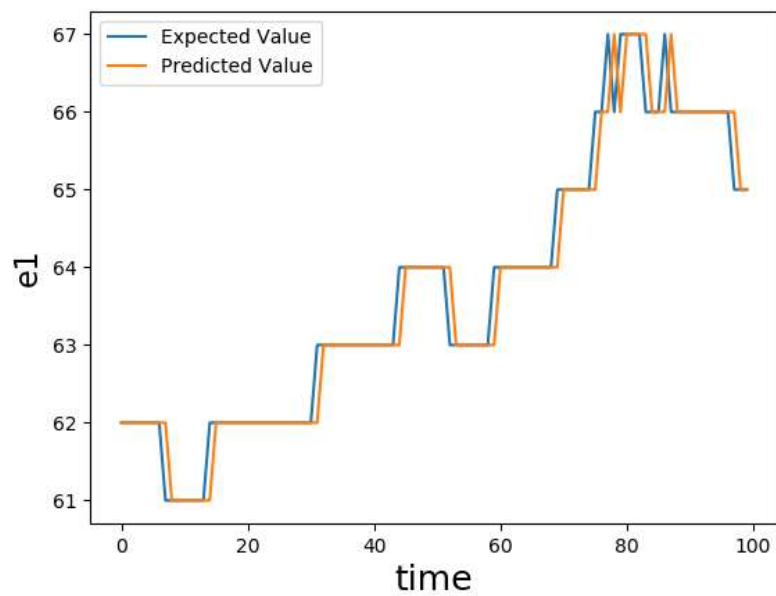




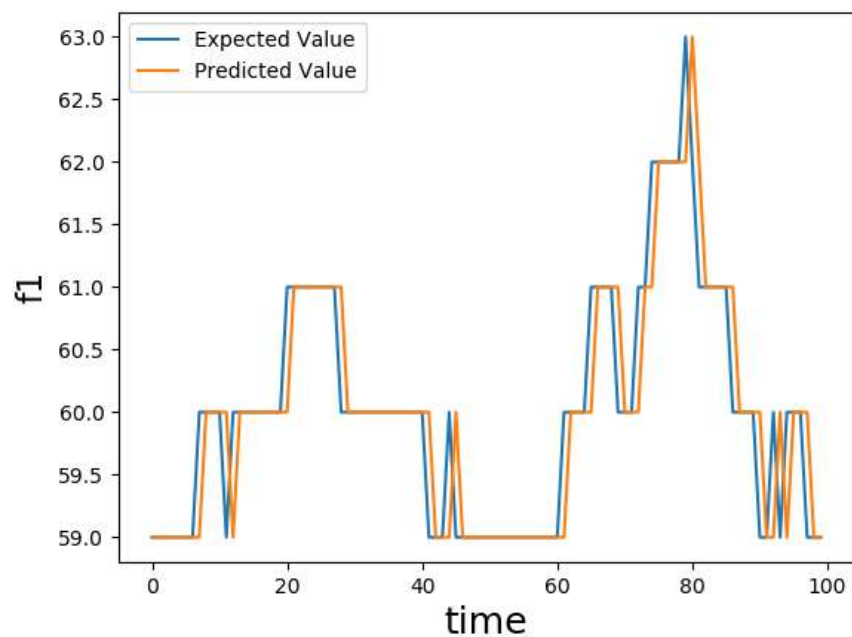
تصویر ۱۰۰ تا از داده های تست فایل D\_ticker برای baseline :



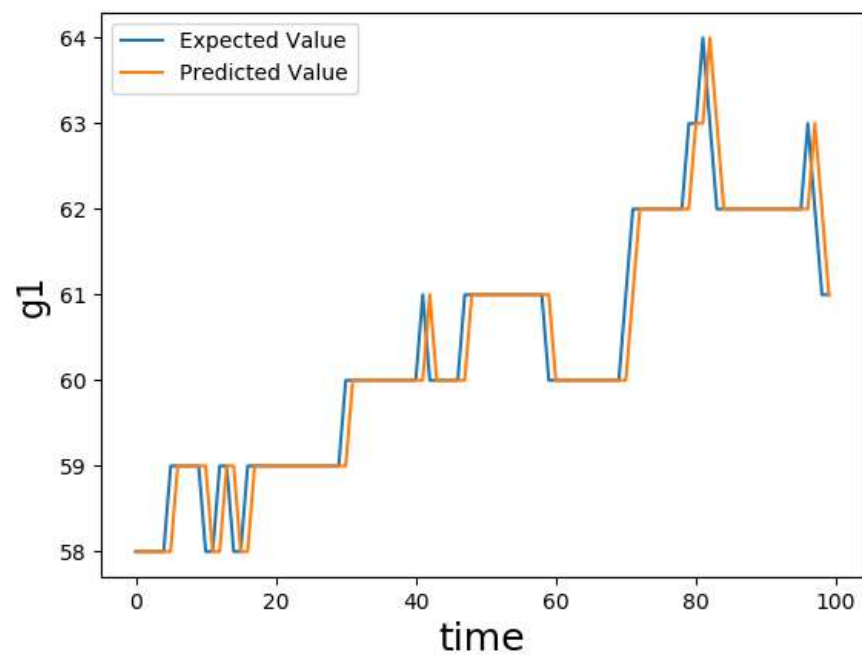
تصویر ۱۰۰ تا از داده های تست فایل E\_ticker برای baseline :



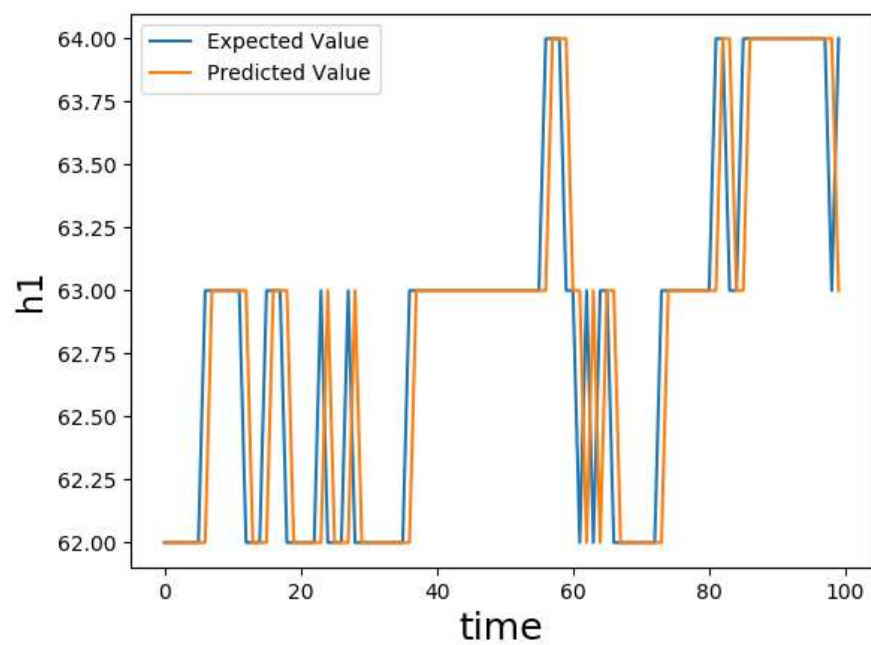
تصویر ۱۰۰ تا از داده های تست فایل F\_ticker برای baseline :



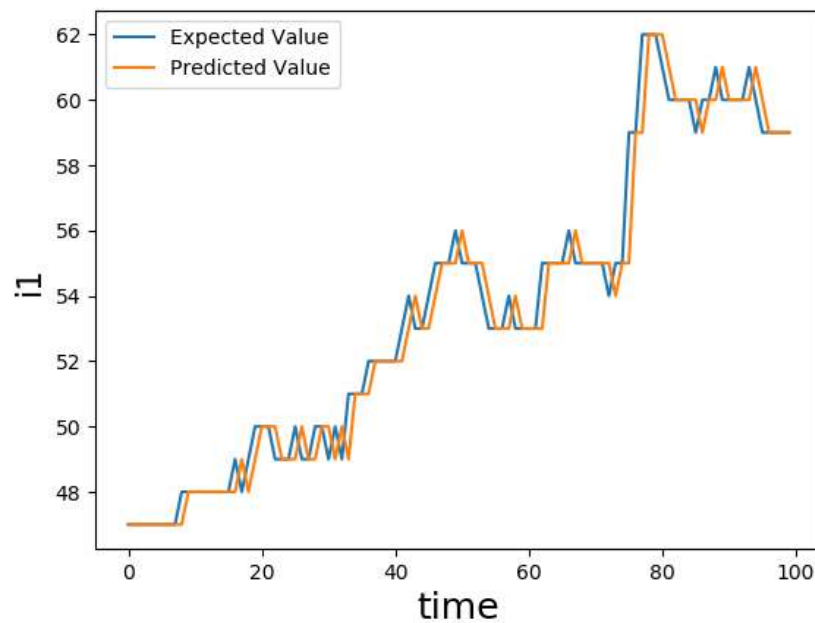
تصویر ۱۰۰ تا از داده های تست فایل G\_ticker برای baseline :



تصویر ۱۰۰ تا از داده های تست فایل H\_ticker برای baseline :



تصویر ۱۰۰ تا از داده های تست فایل I\_ticker برای baseline :



میانگین خطای RMSE و واریانس بعد از ۱۰ بار اجرای مدل baseline روی فایل a\_ticker تا i\_ticker :

I_TICKER	H_TICKER	G_TICKER	F_TICKER	E_TICKER	D_TICKER	C_TICKER	B_TICKER	A_TICKER	
0.733	0.551	0.525	0.588	0.582	0.623	0.634	0.664	0.625	RMSE
0.0001	0.0002	0.0001	0.0001	0.0000	0.0001	0.0000	0.0002	0.0002	VAR

همچنین از داده های smooth شده که در قسمت پردازش داده توضیح داده شد نیز استفاده کردیم ولی به دلیل این که تغییر چندانی در خطاها ایجاد نکرد کارایی زیادی نداشت. نکته ی قابل توجه این است که فقط داده های بخش train را هموار کردیم و برای تخمین از داده های هموار نشده ی test استفاده کردیم.

## بخش سوم :

پیاده سازی مدل svm با کرنل rbf :

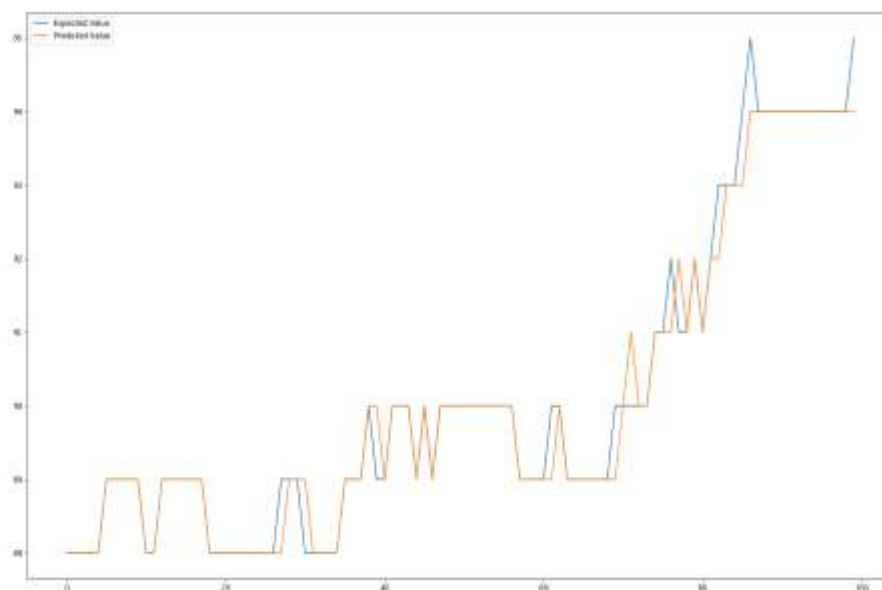
بعد از به دست آمدن مدل اولیه از یک الگوریتم پیچیده تر به نام svm برای ایجاد مدل و تخمین داده ها استفاده کردیم .

همانطور که در بخش قبلی هم گفته شد به طور کلی برای اجرای این الگوریتم ها و پیدا کردن مدلی برای تخمین قیمت آینده نیاز به ایجاد داده های train و test داشتیم که برای این کار بعد از خواندن اطلاعات فایل اول به صورت رندوم ۲۰ درصد از داده ها به عنوان test و ۸۰ درصد دیگر به عنوان train انتخاب شدند.

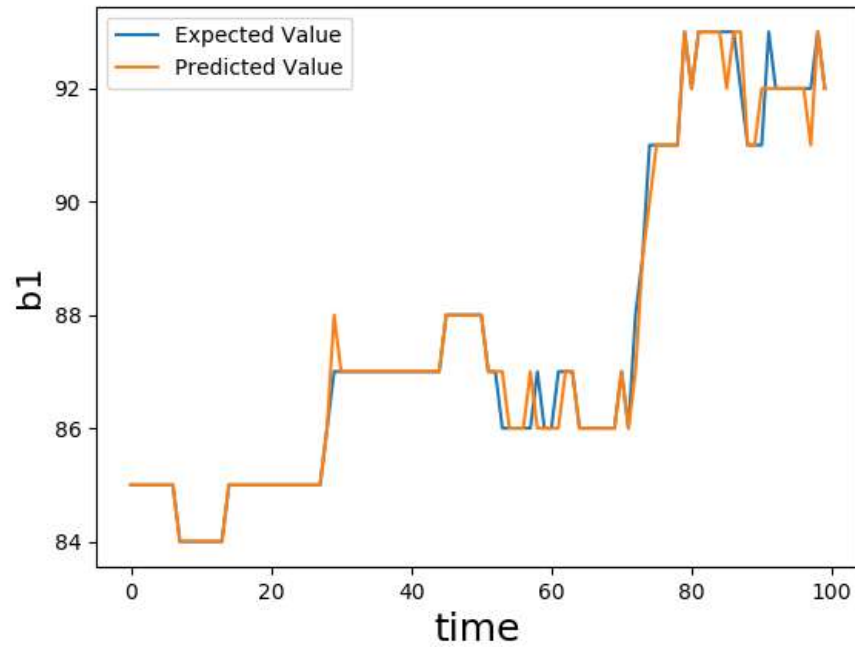
حال به بررسی شکل داده های تخمین زده شده و خطای مربوط به الگوریتم های پیاده سازی شده می پردازیم :

الگوریتم svm با کرنل rbf :

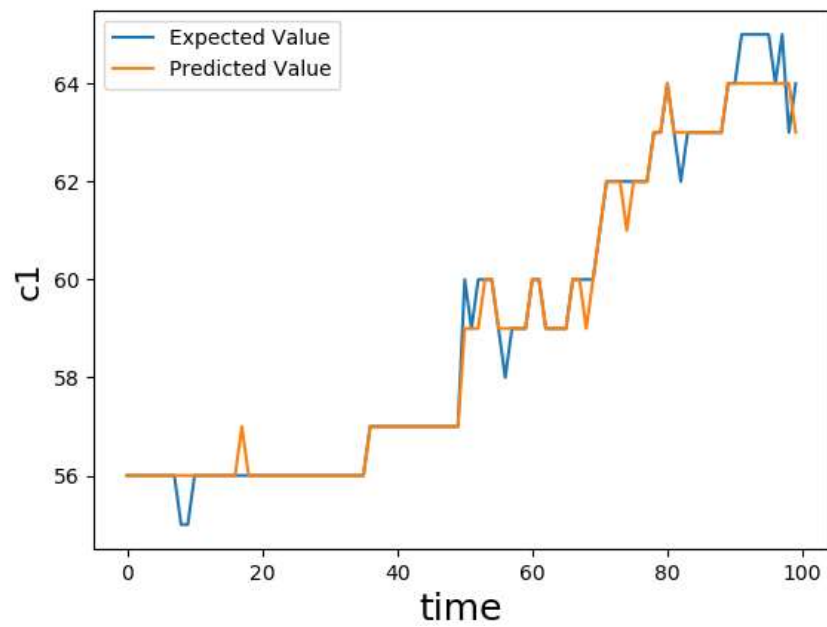
تصویر ۱۰۰ تا از داده های تست فایل A\_ticker برای svm :



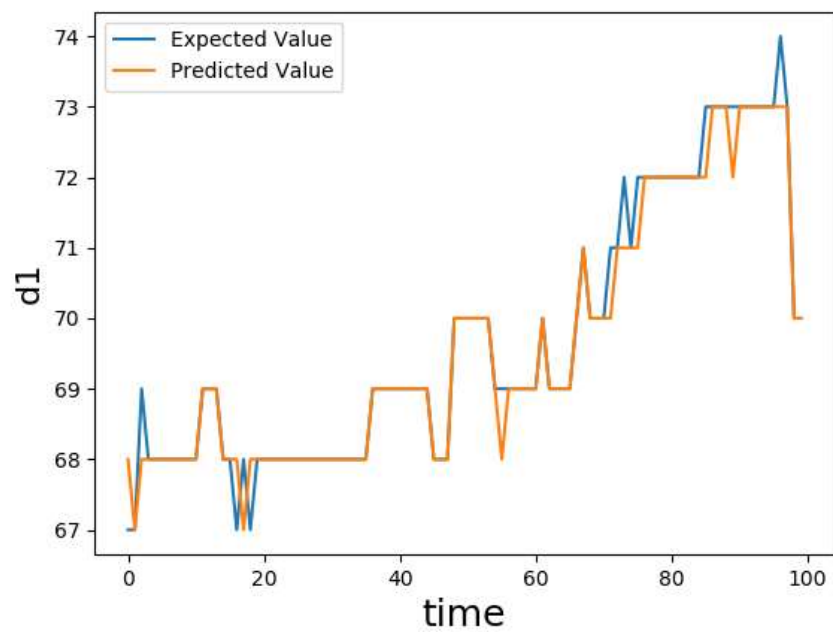
تصویر ۱۰۰ تا از داده های تست فایل B\_ticker برای svm :



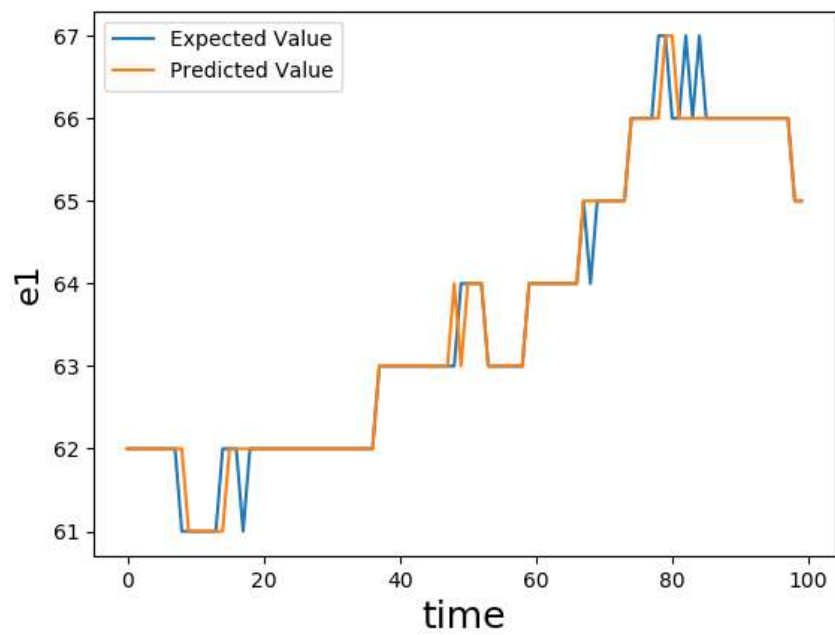
تصویر ۱۰۰ تا از داده های تست فایل C\_ticker برای svm :



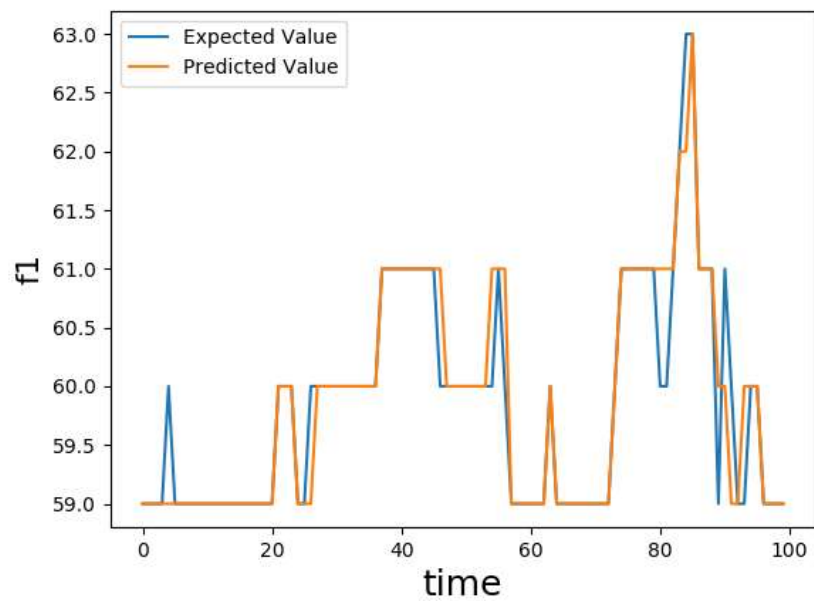
تصویر ۱۰۰ تا از داده های تست فایل D\_ticker برای svm :



تصویر ۱۰۰ تا از داده های تست فایل E\_ticker برای svm :

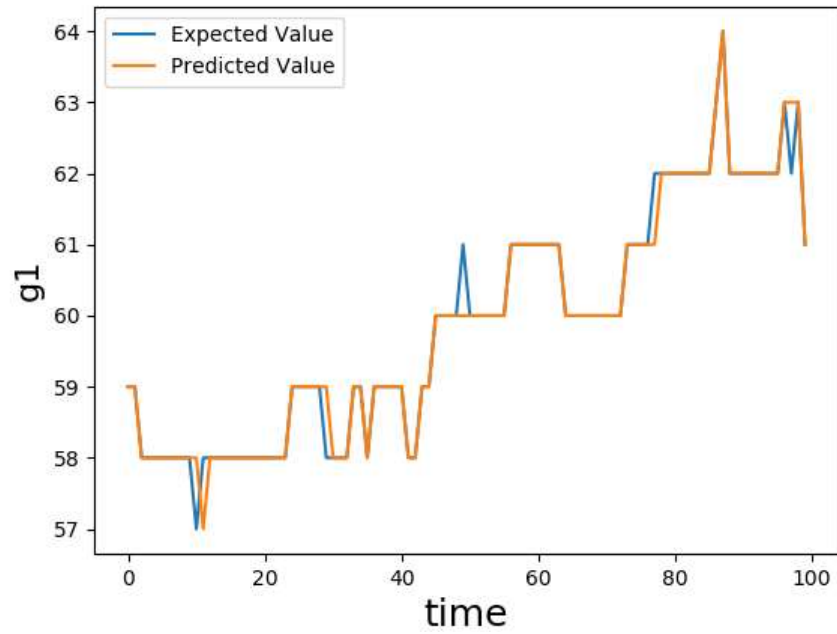


تصویر ۱۰۰ تا از داده های تست فایل F\_ticker برای svm :

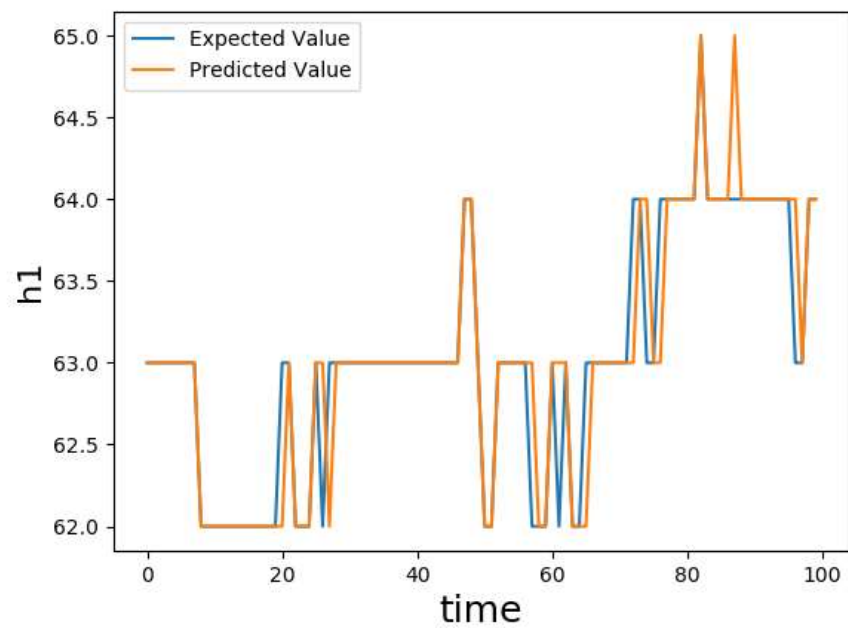




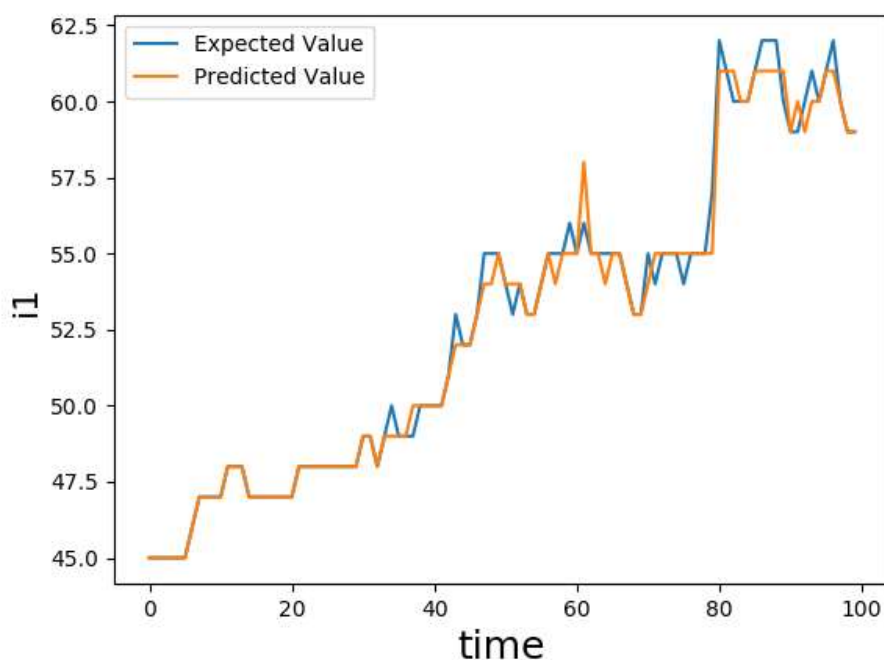
تصویر ۱۰۰ تا از داده های تست فایل G\_ticker برای svm :



تصویر ۱۰۰ تا از داده های تست فایل H\_ticker برای svm :



تصویر ۱۰۰ تا از داده های تست فایل I\_ticker برای svm :



میانگین خطای RMSE و واریانس بعد از ۱۰ بار اجرای مدل SVM روی فایل a\_ticker تا i\_ticker :

I_TICKER	H_TICKER	G_TICKER	F_TICKER	E_TICKER	D_TICKER	C_TICKER	B_TICKER	A_TICKER	
0.323	0.325	0.312	0.361	0.344	0.355	0.303	0.436	0.370	RMSE
0.0000	0.0008	0.0014	0.0022	0.0065	0.0001	0.0001	0.0020	0.0000	VAR

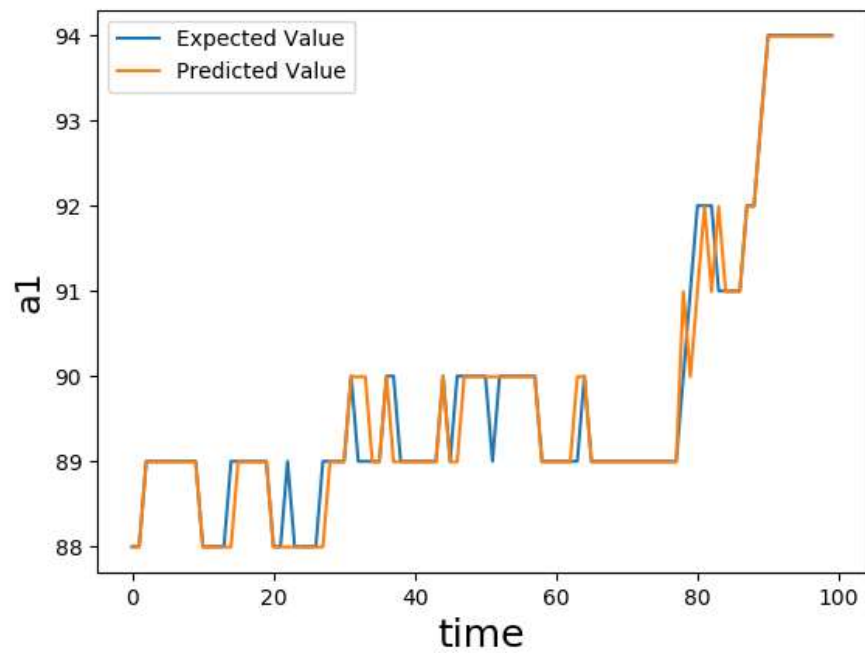
همچنین از داده های smooth شده که در قسمت پردازش داده توضیح داده شد نیز استفاده کردیم ولی به دلیل این که تغییر چندانی در خطاها ایجاد نکرد کارای زیادی نداشت.

## بخش چهارم :

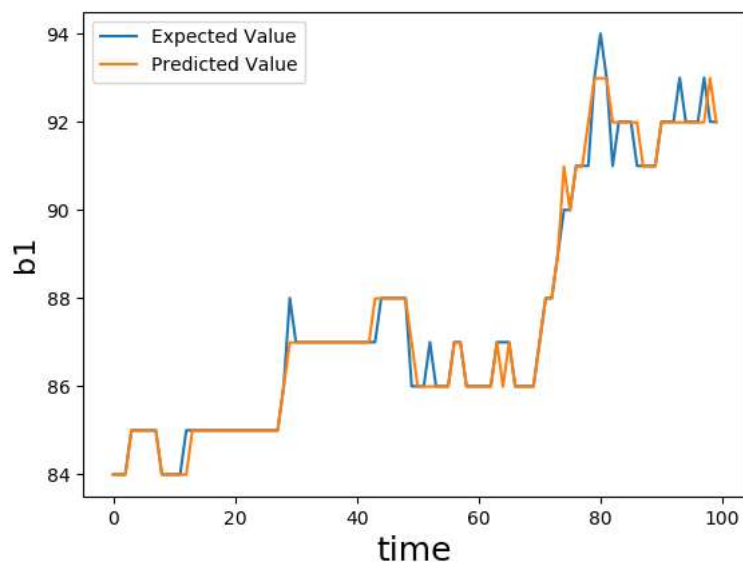
مدل خطی :

توضیحات نمودار چیدمان

تصویر ۱۰۰ تا از داده های تست فایل A\_ticker برای svm :



تصویر ۱۰۰ تا از داده های تست فایل B\_ticker برای svm :



شکل ها برای همه داده ها کشیده شده و بررسی شده ولی برا کوتاه کردن گزارش شکل دو جنس اول را در گزارش قرار داده ایم .

میانگین خطای RMSE و واریانس بعد از ۱۰ بار اجرای مدل LINEAR روی فایل a\_ticker تا i\_ticker :

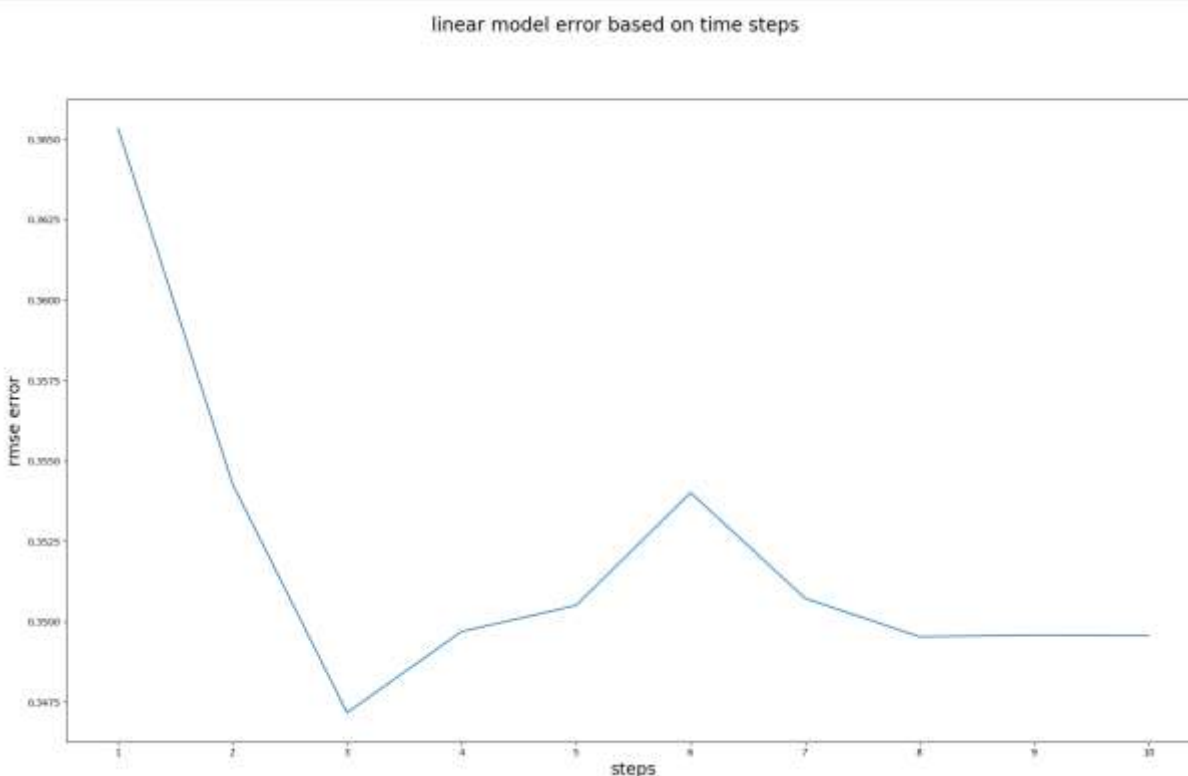
I_TICKER	H_TICKER	G_TICKER	F_TICKER	E_TICKER	D_TICKER	C_TICKER	B_TICKER	A_TICKER	
0.386	0.377	0.387	0.430	0.368	0.401	0.304	0.436	0.370	RMSE
0.0369	0.0170	0.0212	0.0232	0.0167	0.0033	0.0001	0.0019	0.0000	VAR

## بخش پنجم :

استفاده از چند گام زمانی قبلتر برای تخمین :

همانطور که در بالاتر هم گفته شد برای تخمین از مقدار گام قبلی استفاده کردیم در این بخش از چند گام زمانی قبلی برای تخمین نیز استفاده کردیم تا تفاوت خطاها را بررسی کنیم .

- نمودار خطای مدل خطی بر حسب تعداد گام در نظر گرفته شده:



این نمودار میانگین خطای مدل خطی به ازای مقدار گام های در نظر گرفته شده را در ۱۰ اجرای متوالی نشان می دهد. همان طور که مشاهده می شود نقطه کمینه این نمودار در ۳ گام رخ داده از همین رو برای مدل خطی بهینه این است که تا ۳ گام قبل را در نظر بگیریم. البته برای بررسی اینکه این تفاوت معنی دار است یا نه با استفاده از T test این مسئله را بررسی می کنیم:

تعداد دفعات اجرا	واریانس	میانگین	تعداد گام زمانی
10	0.0002	0.3653	1
10	0.0002	0.3472	3

به کمک ابزار موجود در این سایت تست ttest انجام شد و نتیجه این بود که این تفاوت معنی دار است.

<https://www.graphpad.com/quickcalcs/ttest1/?Format=SEM>

Review your data:

Group	Group One	Group Two
Mean	0.347200	0.365300
SD	0.000632	0.000632
SEM	0.000200	0.000200
N	10	10

## Unpaired *t* test results

### P value and statistical significance:

The two-tailed P value is less than 0.0001

By conventional criteria, this difference is considered to be extremely statistically significant.

### Confidence interval:

The mean of Group One minus Group Two equals -0.018100

95% confidence interval of this difference: From -0.018694 to -0.017506

### Intermediate values used in calculations:

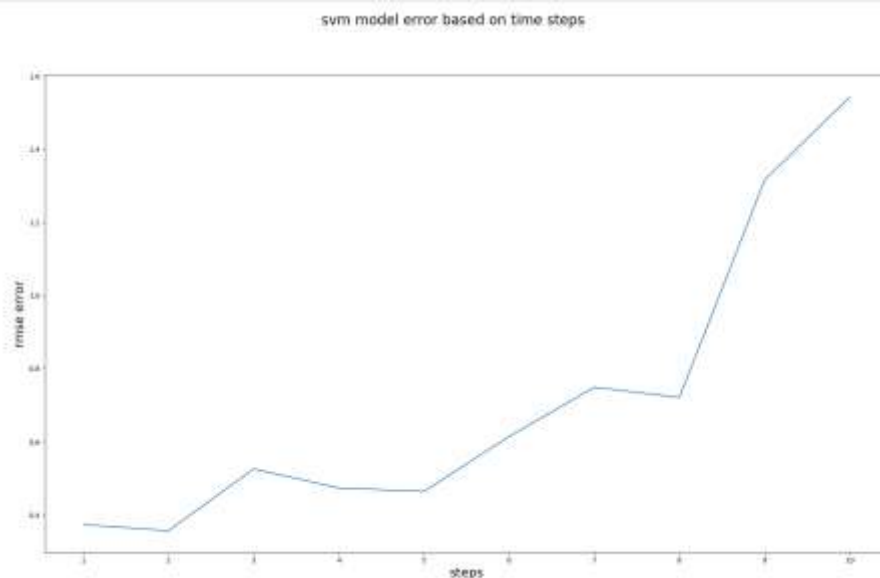
$t = 63.9932$

$df = 18$

standard error of difference = 0.000

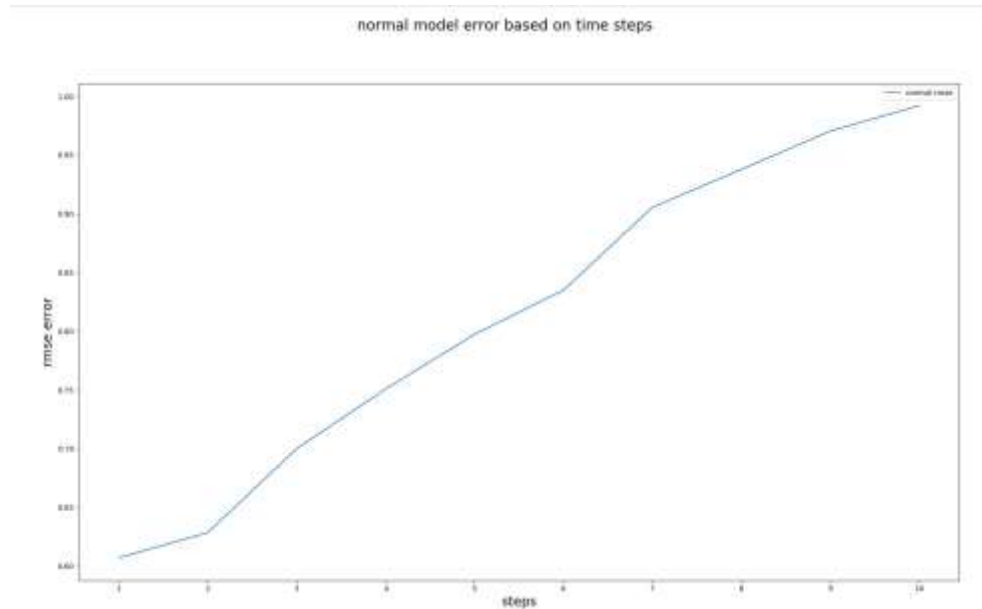
برای همین در مدل خطی تا ۳ گام قبل را در نظر خواهیم گرفت.

- نمودار خطای مدل svm بر حسب تعداد گام در نظر گرفته شده:

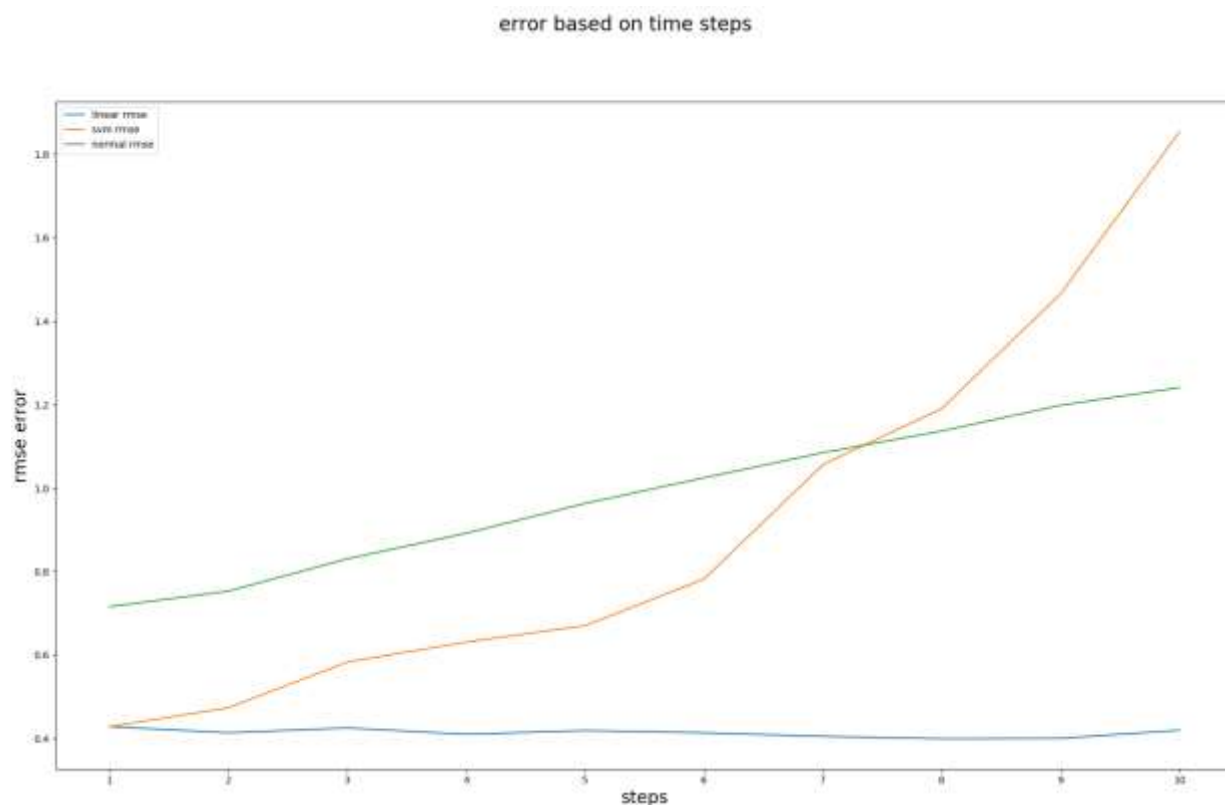


این نمودار نیز همانند نمودار مدل قبل کشیده شده. برای این مدل مشاهده می شود که با افزایش تعداد گام خطا نیز افزایش می یابد.

- نمودار خطای مدل نرمال بر حسب تعداد گام در نظر گرفته شده:
- برای مدل نرمال هم این آزمایش انجام شد. به این معنی که مقدار قیمت در هر گام برابر با میانگین  $k$  گام قبلی قرار می گرفت (که  $k$  تعداد گام است که مورد آزمایش قرار گرفته است).



برای این مدل نیز مشاهده می شود که با افزایش تعداد گام خطا افزایش می یابد.  
همان برای داده B\_ticker:



به طور مشابه می توان میزان خطا بر حسب تعداد گام برای مدل های مختلف را روی همه ارز ها امتحان کرد و بهترین تعداد گام برای هر مدل را پیدا کرد.



## بخش ششم :

بخش نهایی :

نتیجه گیری:

- برای تخمین زدن قیمت طلا کارهای مختلفی انجام شد و ۳ مدل خطی، svm و نرمال امتحان شد. همین طور این موضوع که در نظر گرفتن چند گام قبل خطا را کمتر می کند نیز بررسی شد. تاثیر هموار سازی داده ها در خطا نیز بررسی شد و مشاهده شد که اثر گذار نیست.
- مدل پیش بینی قیمت به صورت کلی به حالت زیر در آمد (برای هر کدام از ارز ها به صورت جداگانه):
- (۱) در نظر گرفتن هر ۳ مدل خطی، svm و نرمال و یافتن بهترین تعداد گام برای هر کدام از این مدل ها به صورت اتومات
  - (۲) اجرای هر کدام از مدل ها با بهترین تعداد گام محاسبه شده مربوط به خودشان از مرحله قبل و محاسبه خطا
  - (۳) انتخاب مدل با خطای کمتر
  - (۴) پیش بینی قیمت بر اساس آن مدل و تعداد گام مربوطه