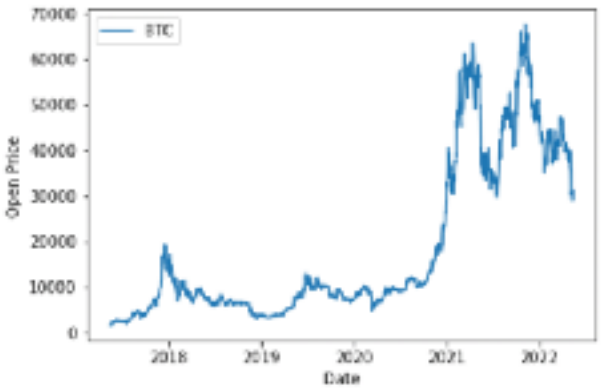
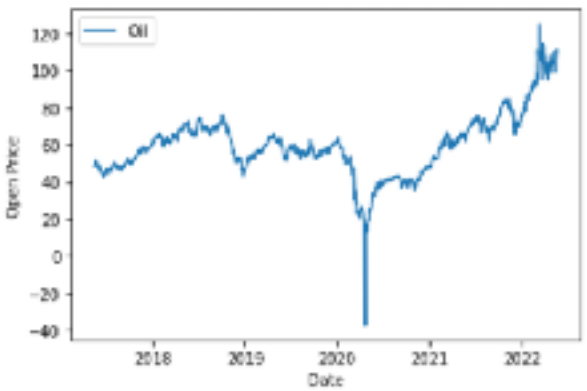
۰) داده

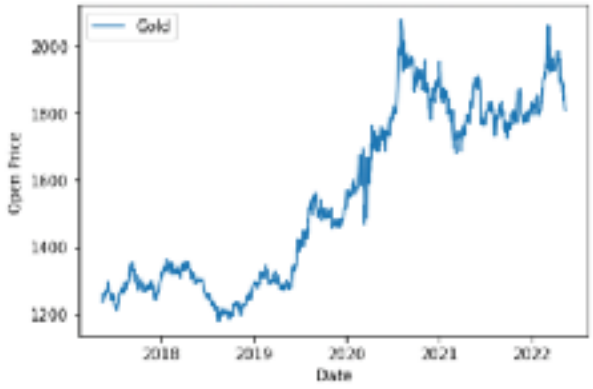
داده‌های قیمت روزانه‌ی بیتکوین را از لینک 1 دانلود کردم. این داده‌ها قیمت close و open هر روز از سال 2017 تا 2022 (1826 روز) را شامل می‌شوند. از این دو نرخ بازدهی روزانه (RoR) محسابه شد. شکل دیتافریم نهاییی پس از حذف ستون‌های اضافی و اضافه کردن ستون ROR به صورت زیر است:

	Date	Close	Open	RoR
0	2017-05-17	1801.3	1729.3	0.041635
1	2017-05-18	1881.0	1801.3	0.044246
2	2017-05-19	1962.0	1881.0	0.043062
3	2017-05-20	2040.2	1962.0	0.039857
4	2017-05-21	2044.2	2040.2	0.001961

داده‌های قیمت روزانه‌ی نفت و طلا را نیز به ترتیب از لینک‌های 2 و 3 دانلود کردم و به طور مشابه با داده‌ها بیتکوین، ستون RoR را تشکیل دادم. این داده‌ها نیز برای پنج سال متوالی از 2017 تا 2022 (1260 روز) هستند و تنها تفاوت آن‌ها در این است که قیمتی برای آن‌ها در روزهای تعطیل (شنبه و یکشنبه) ثبت نشده است. نمودارهای زیر تغییرات قیمت open هر یک از این دیتافریم‌ها را برحسب زمان نشان می‌دهد:

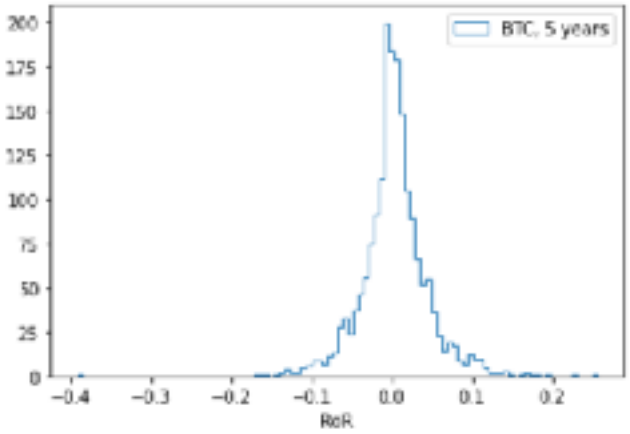






۱) روش‌های کلاسترینگ

نمودار زیر توزیع نرخ‌های بازدهی روزانه برای بیتکوین را در طول پنج سال نشان می‌دهد. هماطور که مشخص است اکثر روزها داری RoR نزدیک به صفر هستند.

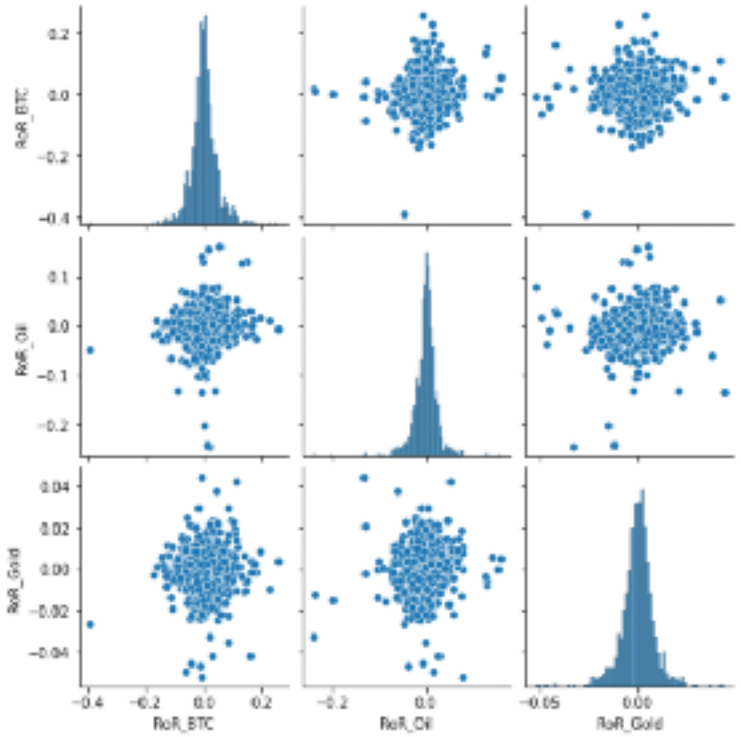


با توجه به لینک 4 نرخ بازدهی نزدیک به 5% یک RoR مناسب معرفی شده است. درنتیجه برای کلاس بندی داده‌ها به صورت زیر عمل کردم:

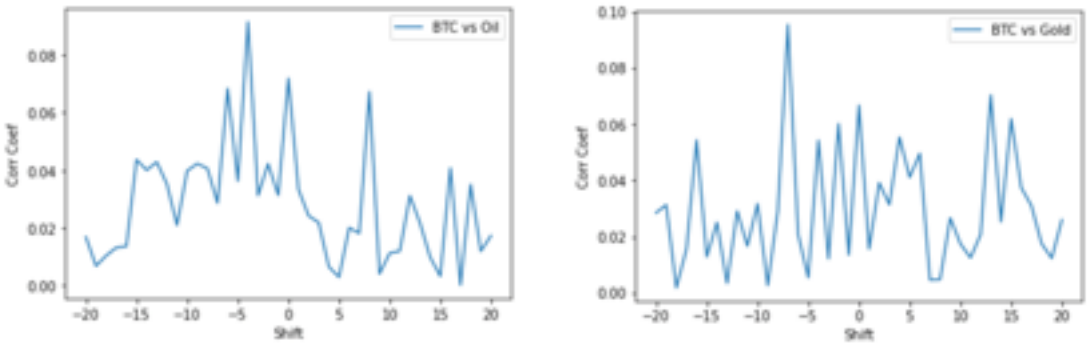
RoR	Label	Count
< -0.05	-1	154
> 0.05	1	169
else	0	1503
	sum	1826

۲) تحلیل داده

نمودار زیر نرخ‌های بازدهی بیتکوین، نفت و طلا را برحسب یکدیگر نمایش می‌دهد. (تابع توزیع هریک به صورت جداگانه نیز نمایش داده شده است)



با توجه به نمودارهای بالا نمی‌توان انتظار همبستگی بالایی برای نرخ بازدهی بیتکوین برحسب نفت یا طلا داشت. با اعمال لگ زمانی در بازه‌ی ±20 روز و محاسبه‌ی ضریب همبستگی (Corr Coef) بیتکوین-نفت و بیتکوین-طلا می‌توان اثر لگ زمانی را بررسی کرد که در دو نمودار زیر رسم شده‌اند.

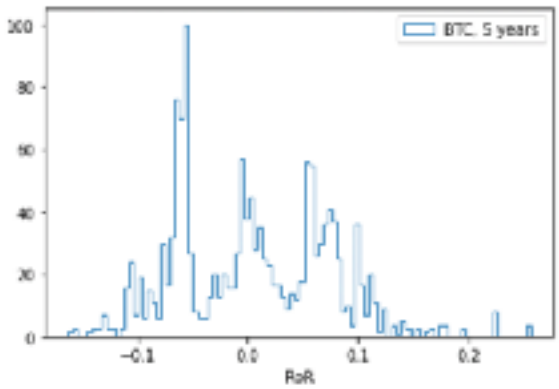


با توجه به نمودارها، بیشینه همبستگی نرخ بازدهی بیتکوین با نرخ بازدهی نفت پس از 5 روز و با نرخ طلا پس از 7 روز است. البته ضریب همبستگی در هردوی این بیشنه‌ها تقریبا برابر با 0.1 است که ضریب همبستگی پایینی محسوب می‌شود.

۳) یادگیری ماشین

با توجه به توزیع نامتوازن داده‌ها در سه کلاس، در صورتی که همه داده‌ها در کلاس صفر پیش‌بینی شوند، دقت بالایی حاصل خواهد شد. ولی کلاس‌های 1 و 1- اهمیت بیشتری برای پیش‌بینی دارند. برای حل کردن این مشکل دو روش متفاوت را امتحان کردم.

۱- ریسمپلینگ (resampling):
در این روش با استفاده از کتابخانه‌ی sklearn داده‌ها کلاس‌های 1 و 1- را افزایش و داده‌های کلاس 0 را کاهش دادم به طوری در هر کلاس تعدا برابر 485 داده موجود باشد. نمودار توزیع RoR ها بعد از انجام ریسمپلینگ به صورت زیر درآمد:



البته باید توجه داشت که داده‌ی تست از قبل جدا شده باشد و تحت ریسمپلینگ قرار نگیرد. بعد از آموزش داده‌های train با استفاده از مدل SVM، نتایج زیر بر روی داده‌ها test بدست آمد:

Classification report :

	precision	recall	f1-score	support
-1	0.17	0.32	0.22	34
0	0.86	0.61	0.71	302
1	0.08	0.24	0.12	29
accuracy			0.55	365
macro avg	0.37	0.39	0.35	365
weighted avg	0.73	0.55	0.62	365

دقت کل برای این مدل 55 درصد است و به ترتیب 32 و 24 درصد از داده‌های کلاس 1- و 1 به درستی پیش‌بینی شده‌اند.

۲- افزایش وزن خطا برای داده‌های کلاس 1 و -1:
در این روش با توجه به اهمیت کلاس، خطای بزرگتری برای پیش‌بینی اشتباه درنظر می‌گیریم. به طوری ماشین به سمت یادگرفتن کلاس‌های کم جمعیت‌تر برود. نسبت وزن خطا برای کلاس‌های 1- ، 1 و 0 به ترتیب 100، 10و 1 را درنظر گرفتم و پس از آموزش یک شبکه عصبی، نتایج زیر برای داده‌های test بدست آمد:

Classification report :

	precision	recall	f1-score	support
1	1.00	0.02	0.05	42
0	0.82	0.44	0.58	366
-1	0.08	0.57	0.14	35
accuracy			0.41	443
macro avg	0.63	0.35	0.25	443
weighted avg	0.78	0.41	0.49	443

همان طور که انتظار می‌رود داده‌های کلاس 1- با دقت بالاتری نسبت به قبل (57 درصد) پیش بینی شده‌اند ولی دقت پایین کلاس 1 با توجه به اینکه وزن بالاتری نسبت به کلاس 0 دارد، به این معناست که برای ماشین تمایز کلاس 1 از 1- دشوار است.
مشخصات شبکه عصبی استفاده در جدول زیر آمده است:

Layer (type)	Output Shape	Param #
flatten_19 (Flatten)	(None, 4)	0
dense_75 (Dense)	(None, 8)	40
dense_76 (Dense)	(None, 1)	9

Total params: 49
Trainable params: 49
Non-trainable params: 0

این شبکه عصبی 49 پارامتر مختلف دارد که با توجه به تعداد کم داده‌های train برای جلوگیری از overfit انتخاب شده است.
همچنین دقت کل این شبکه عصبی برابر با 0.41 درصد است که کمتر از روش اول می‌باشد ولی داده‌های کلاس 1- که اهمیت بیشتری دارند، با دقت بالاتری پیش‌بینی شده‌اند.