به نام خدا

عنوان پروژه : تعیین روند رفتار سهام های بورس ایران با استفاده از الگوریتم های یادگیری ماشین

تاریخ:۱۴۰۰/۰۷/۰۷

نوع مسئله:یادگیری بدون نظارت خوشه بندی،یادگیری با نظارت رگرسیون

مجموعه داده:قیمت سهام ایران از تاریخ۲۰۰۵/۰۱/۰۱تا۲۰۲۰/۰۶/۰۱

چکیده:

استفاده از تکنولوژی و فناوری‌های به روز همیشه می‌تواند ما را یک قدم جلوتر از دیگران قرار دهد، به ویژه وقتی صحبت از زمینه‌هایی مثل بورس و سرمایه گذاری باشد. دو نمونه از بهترین تکنولوژی‌های امروز در بازار بورس هوش مصنوعی و یادگیری ماشین نام دارند که به تازگی پا به این عرصه گذاشته‌اند.با استفاده از این دو عامل و پلتفرم‌های وابسته به آن می‌توان بازدهی معاملات را افزایش داد، در هزینه و وقت صرفه جویی کرد و نهایتاً درصد ریسک را تا حد قابل توجهی کاهش داد.در سال های اخیر استفاده از هوش مصنوعی و یادگیری ماشین یکی از امیدوار کننده ترین ابزار ها برای تجزیه و تحلیل و پیش بینی داده های سری زمانی که شامل داده های بازار های مالی می شود بوده است، پیش بینی بازار سهام عملیاتی است که تلاش می کند ارزش آینده سهام یک شرکت یا سایر نماد های مالی معامله شده در بازار های مالی را تعیین کند(پیش بینی کند).پیش بینی موفقیت آمیز قیمت آینده سهام ، سود سرمایه گذار را به حداکثر می

رساند در این پژوهش ما الگوریتم های مختلف یادگیری ماشین را روی داده ها بورس ایران آموزش داده و سپس دقت آنها را مورد ارزیابی قرار داده ایم روش کار پیشنهادی ما در این پژوهش به این شرح می باشد که ابتدا با استفاده از الگوریتم های خوشه بندی نمادهای را که رفتار مشابه بهم دارند را در یک خوشه قرار می دهیم سپس با استفاده از الگوریتم های رگرسیون سعی میکنیم قیمت سهام ها را از روی قیمت سایر سهام های موجود در آن خوشه پیش بینی کنیم، با استفاده از این روش دقت های زیر بدست امده است:برای الگوریتم درخت تصمیم دقت ۹۹,۹٤ روی داده های ترین و دقت ۹۹,۸۹ روی داده های تست با میزان خطای ۳۸۸,۸۳۸ و الگوریتم لاسو با دقت ۹۹,۹۰ روی داده های ترین و دقت۹۹,۹۲ روی داده های تست با میزان خطای۳۳۳۸,۳٤٦ با معیار مربعات خطا محاسبه گردیده است.

فواید پیش بینی قیمت های بازار مالی با استفاده از الگوریتم های هوش مصنوعی چیست؟

در عصری زندگی می‌کنیم که تکنولوژی تا ریزترین بخش‌های زندگی فردی و اجتماعی انسان وارد شده است و اجتناب از آن امکان‌پذیر نیست. از جمله بازارهایی که چند سالی می‌شود به اجتناب‌ناپذیر بودن این حقیقت پی برده‌اند بازارهای مالی هستند. ورود بازار سرمایه به عصر تکنولوژی با معاملات الگوریتمی اتفاق افتاد،همه افراد از متضرر شدن در بازار بیزارند. آگاهی از بستر ایمن و مطمئن سرمایه‌گذاری برای افراد ضروری است. الگوریتم های یادگیری ماشین با استفاده از داده های بسیار زیادی که متخصصین در

اختیارش قرار میدهند آموزش می بیند. سپس با توجه به یک چهارچوب منطقی به شما می‌گوید که به عنوان مثال این شرکت سابقه‌ی خوبی دارد یا نه. این مورد حتی در جهت شناسایی کلاهبرداران نیز مورد استفاده قرار می‌گیرد که شما را از معرض مورد تقلب قرار گرفتن ایمن می‌سازد.یکی از مهم‌ترین ویژگی‌های هوش مصنوعی و یادگیری ماشین این است که نیازی به مداخله انسانی ندارد. این موضوع سرعت ما را در زندگی روزمره به شدت افزایش می‌دهد، مدیریت زمان ما را معنادارتر و دیدگاه ما را بزرگ‌تر میکند. به بیان بهتر به کارگیری هوش مصنوعی و یادگیری ماشین مانند استخدام کارمندی است که کاملا دقیق بوده، به صورت مستمر درحال یادگیری است، خسته نمی‌شود، قدرت پردازش خوبی دارد و می‌تواند سفارشی سازی شود. آیا واقعاً انسانی به این شکل وجود دارد؟ مسلماً خیر. سیستم های مبتنی بر هوش مصنوعی و یادگیری ماشین کاملا بر اساس منطق تصمیم گیری می کند و فاقد هرگونه احساسی می باشند از این رو بدون درنظر گرفتن خرافات و احساسات به ارائه‌ی نتایج از داده‌ها زمان دار که به دقت آن می‌افزاید، اقدام می‌کند، استفاده از هوش مصنوعی سبب کاهش ریسک و صرفه جویی در زمان می شود.

اهداف:

۱-شناسایسی سهام های که رفتار مشابه دارند

۲-پیش بینی قیمت سهام ها و ارزیابی عملکرد الگوریتم های یادگیری ماشین

۳-افزایش بازدهی معاملات

تعاریف:

سهام:سهام یک ورق بهادار است که بر مالکیت بخشی از یک شرکت دلالت دارد. این سهام به دارنده‌اش حق می‌دهد که متناسب با تعداد آن سهام در بخشی از دارایی‌های شرکت و سود آن شریک باشد. واحدهای سهام سهم نامیده می‌شوند.سهام به طور عمده در بورس اوراق بهادار خرید و فروش می‌شود، هرچند می‌توان سهام را به صورت خصوصی نیز معامله کرد و پایه و اساس بسیاری از اوراق بهادار سرمایه‌گذاران شخصی است. این معاملات باید مطابق مقررات دولت باشد که به منظور محافظت از سرمایه‌گذاران در برابر اقدامات مخرب و کلاهبرداری تدوین شده‌اند. از نظر تاریخی، در طولانی مدت، سهام از سایر سرمایه‌گذاری‌ها بهتر بوده است. سهام را می‌توان از اکثر کارگزاری‌ها به صورت آنلاین خرید و فروش کرد.

هوش مصنوعی:[1]هوش مصنوعی که گاهی اوقات هوش ماشینی نیز نامیده می شود شبیه سازی فرایندهای هوش طبیعی[2]توسط ماشین ها به ویژه سیستم های رایانه ای است. به عبارت دیگر، هوش مصنوعی به سامانه‌هایی گفته می‌شود که می‌توانند واکنش‌هایی مشابه رفتارهای هوشمند انسانی از جمله، درک شرایط پیچیده، شبیه‌سازی فرایندهای تفکری و شیوه‌های استدلالی انسانی و پاسخ موفق به آن‌ها، یادگیری و توانایی کسب دانش و استدلال برای حل مسائل را داشته باشند،بطور خلاصه هوش مصنوعی را دانش ساخت وطراحی عامل هوشمند تعریف کرده اند.این علم کاربرد های فراوانی در علوم رایانه، علوم مهندسی،تجارت،پزشکی و بسیاری از علوم دیگر دارد بعنوان مثال: در پزشکی

[1] Artificial Intelligence
[2] Natural Intelligence

تجزیه و تحلیل صدا قلب،ربات های پرستار،ارائه مشاوره و پیش‌بینی احتمال مرگ بیمار برای هر روش جراحی....،در امورمالی و تجارت تجزیه وتحلیل بازار های مالی،پیش بینی قیمت سهام ها،معاملات الگوریتمی،مدیریت دارای و...از کاربرد های هوش مصنوعی در این علوم هستند. هوش مصنوعی، موضوعی بسیار گسترده است که شاخه‌های متعددی دارد. شاخه های هوش مصنوعی عبارتند از:یادگیری ماشینی[۳]،شبکه های عصبی[٤] سیستم های خبره[٥]، پردازش زبان طبیعی [٦] تشخیص گفتار[۷] و بینایی ماشین[۸] و- رباتیک [۹] و منطق فازی [۱۰]است.

یادگیری ماشین:یادگیری ماشینی شاخه‌ای از هوش مصنوعی[۱۱]و علوم کامپیوتر[۱۲]است که بر استفاده از داده‌ها و الگوریتم‌ها برای تقلید از روشی که انسان‌ها یاد می‌گیرند تمرکز دارد و به تدریج دقت آن را بهبود می‌بخشد. یادگیری ماشین به عنوان بخشی از هوش مصنوعی در نظر گرفته می شود که به مطالعه الگوریتم های کامپیوتری می پردازد که می تواند به طور خودکار از طریق تجربه و با استفاده از داده ها بهبود یابد. الگوریتم‌های یادگیری ماشین مدلی را بر اساس داده‌های نمونه می‌سازند که به داده‌های آموزشی معروف است تا پیش‌بینی‌ها یا تصمیم‌گیری‌ها را بدون برنامه‌ریزی صریح انجام دهند.الگوریتم‌های یادگیری ماشین در کاربردهای متنوعی مانند پزشکی، فیلتر کردن

---

[۳] machine learning
[٤] Neural network
[٥] Expert Systems
[٦] Natural language processing
[۷] speech recognition
[۸] Machine vision
[۹] robotic
[۱۰] Fuzzy logic
[۱۱] artificial intelligence
[۱۲] computer science

ایمیل، تشخیص گفتار و بینایی کامپیوتری استفاده می‌شوند.انواع الگوریتم های یادگیری ماشین عبارت اند از:درخت تصمیم[۱۳]،لاجستیک رگرسیون[۱۴]،ماشین بردار پشتیبان[۱۵]،دسته بند بیز[۱۶]،نزدیک ترین همسایه[۱۷]و.... می توان اشاره کرد.

یادگیری عمیق[۱۸]:یادگیری عمیق که در زبان فارسی به یادگیری ژرف نیز ترجمه شده است بخشی از خانواده یادگیری ماشینی می باشد که بر روش های تمرکز دارد که مبتنی بر الگوریتم های شبکه عصبی مصنوعی[۱۹]هستند. این الگوریتم ها د تلاش اند که مغز انسان را شبیه‌سازی کنند.به طور خلاصه در یادگیری عمیق شبکه های عصبی مصنوعی و الگوریتم های مشابه مغز بشر از مجموعه های عظیم داده مهارت های مورد نظر را فرا می گیرند.همانطور که ما از طریق تجربه چیزهای جدید یاد می گیریم الگوریتم یادگیری عمیق نیز با هر بار تکرار یک کار مهارت خود را نسبت به دفعات قبلی بهبود می بخشد. دلیل استفاده از عبارت یادگیری عمیق این است که شبکه های عصبی لایه های مختلف یا عمیقی دارند که یادگیری را ممکن می سازد.

داده کاوی[۲۰]:داده کاوی فرایندی برای تبدیل داده های خام به اطلاعات مفید می باشد،داده کاوی فرآیند استخراج و کشف الگوها در مجموعه داده های بزرگ است که شامل روش هایی در محل تلاقی یادگیری ماشین ، آمار و سیستم های پایگاه داده است.به

---

[۱۳] Decision tree
[۱۴] Logistic Regression
[۱۵] Support vector machine
[۱۶] Naive Bayes Classifiers
[۱۷] KNN
[۱۸] Deep learning
[۱۹] Artificial neural network
[۲۰] Data mining

عبارت دیگر داده کاوی یک زیرشاخه بین رشته ای علوم کامپیوتر و آمار با هدف کلی استخراج اطلاعات (با روشهای هوشمند) از مجموعه داده و تبدیل اطلاعات به یک ساختار قابل درک برای استفاده بیشتر است.

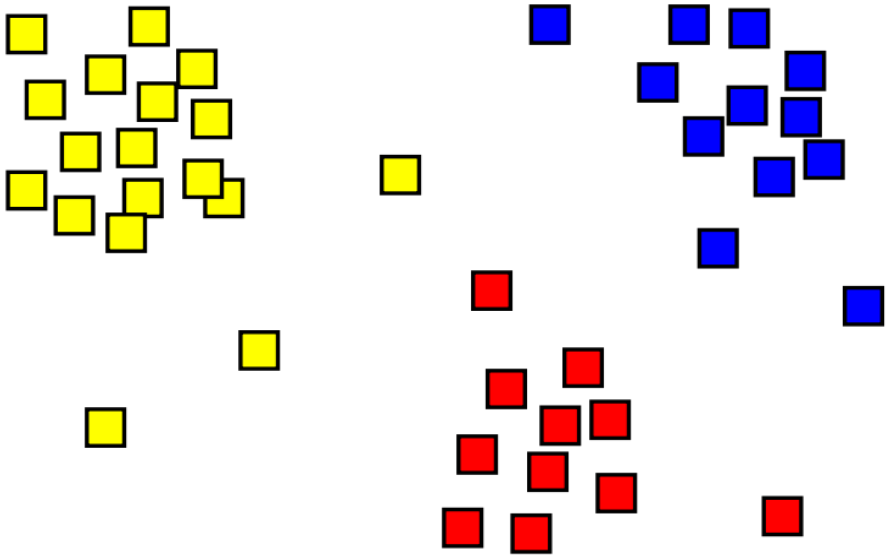
الگوریتم های معاملاتی: استفاده از برنامه‌های کامپیوتری برای ورود به سفارش‌های معاملاتی بدون دخالت انسان. این الگوریتم‌ها که می‌توانند بیش از یکی باشند، برای انجام معاملات بررسی‌های لازم را از جنبه‌های گوناگونی مانند زمان‌بندی، قیمت و حجم روی سفارشات و بازار انجام داده و تصمیم می‌گیرند. این امر کمک می‌کند تا بازار سرمایه به روشی اصولی‌تر و به دور از دخالت احساسات انسانی پیش رود که یکی از نتایج آن بالارفتن نقدینگی در بازار است.

خوشه بندی:[۲۱]تجزیه خوشه ای یا خوشه بندی، وظیفه گروه بندی مجموعه ای از اشیاء به گونه ای است که اشیاء در یک گروه (که خوشه نامیده می شود) بیشتر به یکدیگر شباهت دارند تا اشیاء گروه های دیگر .این یک وظیفه اصلی تجزیه و تحلیل داده های اکتشافی و یک تکنیک رایج برای تجزیه و تحلیل داده های آماری است که در بسیاری از زمینه ها از جمله تشخیص الگو ،[۲۲]تجزیه و تحلیل تصویر ، بازیابی اطلاعات ، بیوانفورماتیک ، فشرده سازی داده ها ، گرافیک رایانه ای و یادگیری ماشین[۲۳]استفاده می شود .خوشه بندی یک روش یادگیری ماشین است که شامل گروه بندی نقاط داده است . با توجه به مجموعه ای از نقاط داده ، می توانیم از یک الگوریتم خوشه بندی برای طبقه

---

[۲۱] clustering
[۲۲] pattern recognition
[۲۳] Machine Learning

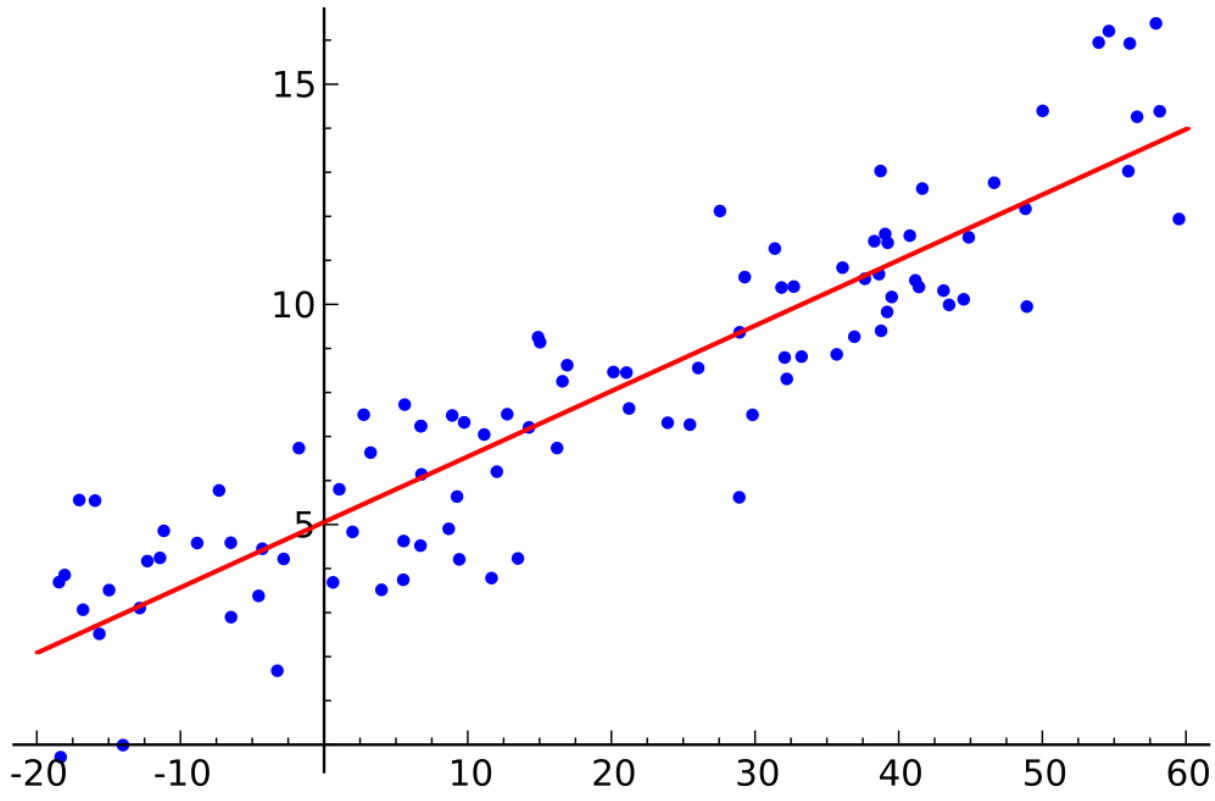
بندی هر نقطه داده در یک گروه خاص استفاده کنیم .از نظر تئوری ، نقاط داده ای که در یک گروه هستند باید دارای ویژگی ها ویا ویژگی های مشابه باشند ، در حالی که نقاط داده در گروه های مختلف باید دارای خواص ویا ویژگی های بسیار متفاوتی باشند . خوشه بندی یک روش یادگیری بدون نظارت[۲]است و یک تکنیک رایج برای تجزیه و تحلیل داده های آماری است که در بسیاری از زمینه ها استفاده می شود .تکنیک های خوشه بندی زمانی که هیچ کلاسی برای پیش بینی وجود نداشته باشد ، و داده ها باید به گروه های طبیعی تقسیم شوند ، کاربرد دارد.



[۲] unsupervised learning

خوشه بند کی مینز:[۲۹] این خوشه بند داده ها را به kخوشه تقسیم بندی می کند و تلاش می کند که واریانس داده های موجود در هر خوشه(واریانس درون خوشه ای) را به حداقل برساند.

رگرسیون: در مدل سازی آماری ، تحلیل رگرسیون مجموعه ای از فرایندهای آماری برای تخمین روابط بین یک متغیر وابسته و یک یا چند متغیر مستقل است. رایج ترین شکل تجزیه و تحلیل رگرسیون ، رگرسیون خطی است که در آن می توان خطی را پیدا کرد که بیشترین تناسب را با داده ها بر اساس یک معیار ریاضی خاص دارد.



[۲۹] K-Means

تحلیل رگرسیون تکنیکی آماری برای بررسی و مدل‌سازی ارتباط بین متغیرها است. رگرسیون تقریباً در هر زمینه‌ای از جمله مهندسی، فیزیک، اقتصاد، مدیریت، علوم زیستی، بیولوژی و علوم اجتماعی برای برآورد و پیش‌بینی مورد نیاز است.

رگرسیون خطی:[۲۶]رگرسیون خطی ساده ترین و پرکاربردترین تکنیک آماری برای مدل سازی پیش بینی مقادیر پیوسته است. این اساساً یک معادله به ما می دهد ، جایی که ما ویژگی های خود را به عنوان متغیرهای مستقل داریم ، که متغیر هدف ما به آن وابسته است.

رگرسیون چندجمله ای:[۲۷]رگرسیون چند جمله ای شکل دیگری از رگرسیون است که در آن حداکثر توان متغیر مستقل بیشتر از ۱ است. در این تکنیک رگرسیون ، بهترین خط مناسب خط مستقیم نیست بلکه به شکل منحنی است.

درخت تصمیم:الگوریتم درخت تصمیم به خانواده الگوریتم های یادگیری ماشین با نظارت[۲۸] تعلق دارد. می توان از این الگوریتم برای حل مسائل طبقه بندی و رگرسیون استفاده کرد. هدف این الگوریتم ایجاد مدلی است که مقدار یک متغیر هدف را پیش بینی می کند، که برای این منظور درخت تصمیم از نمایش درختی برای حل مسئله استفاده می کند. گره برگ مربوط به یک برچسب کلاس است و ویژگی ها در گره داخلی درخت نشان داده می شوند.

---

[۲۶] Linear regression
[۲۷] Polynomial regression
[۲۸] supervised machine learning

روش کار:

از آنجا که ما در این پروژه میخواهیم الگوهای میان داده ها را کشف کنیم باید از الگوریتم های خوشه بندی استفاده کنیم،این الگوریتم ها سهام ها با رفتار های مشابه را در یک خوشه قرار میدهند که ما را به نتیجه که میخواهیم می رسانند،الگوریتم مورد استفاده در این پروژه الگوریتم خوشه بند کی مینز می باشد.چالش اصلی در اموزش الگوریتم نام برده شده تعیین تعداد خوشه ها یا همان K می باشد که برای پیدا کردن بهترین مقدار K دو معیار اینرسی[۲۹]و نمره سیلوئیت[۳۰]استفاده می کنیم.سپس سهام های موجود در یک خوشه را به دوبخش آموزش و ارزیابی تقسیم میکنیم و الگوریتم های رگرسیون را آموزش داده و ارزیابی میکنیم، هدف ما از این کار این است که میزان دقت الگوریتم خوشه بندی را ارزیابی کنیم اگر الگوریتم های رگرسیون دقت خوبی داشته باشند نشان دهنده این است که الگوریتم خوشه بندی خوب عمل کرده است.

شرح گزارش:

ایتدا کتابخونه های پانداز:برای خوندن فایل و تجزیه و تحلیل دیتافریم،نامپای:برای کار با آرایه های نامپای،مت پلات لیب:برای رسم نمودار،سیبورن:برای رسم نمودار،سایکیت لرن برای:استفاده از الگوریتم های یادگیری ماشینی،استاندارد کردن دیتا و اندازه گیری خطا فراخوانی شده اند. این دیتاست شامل قیمت سهام ۷۱۰ شرکت شامل۱۰۲۸۱۷۷رکورد و ۱۰ ستون ویژگی می باشد ،ستون ها تایم[۳۱]و پر[۳۲]به ترتیب

[۲۸]inertia
[۲۹]Silhouette score
[۳۱]TIME
[۳۲]Per

داری مقدار یکتای صفر و دی[۳۳] می باشند و تاثیری بر روی مدل و قیمت سهام ها ندارند به همین علت این دو ستون از دیتاست حذف شدند.تنها رکورد شماره ۸۱۹۲۹۰ با مقادیر نامعلوم[۳۴] در دیتاست وجود دارد که رکورد مورد بحث از دیتاست با روش زیر حذف گردید:

Find the row number containing the NAN value

```
In [12]: data[data.Close.isna()==True]
```

Out[12]:

	Ticker	Per	DTYYYYMMDD	TIME	Open	High	Low	Close	Vol	Openint
819290	TAT_Share	d	20111212	0	NaN	NaN	NaN	NaN	4836900	2768

drop nan and reset index

```
In [13]: data.drop(819290 , axis=0, inplace=True)
         data.reset_index(inplace=True,drop=True)
```

ستون دی تی وای....[۳۵] نشان دهنده تاریخ ثبت قیمت ها می باشد که به صورت یک رشته حروف[۳۶] در دیتاست ذخیر شده است، مقادیر موجود در این ستون را بصورت دیت تایم باشند تا بتوان از آنها استفاده کرد به همین جهت با استفاده از قطعه کد زیر ابتدا ستونی به نام دیت که شامل زمان های ثبت تاریخ با فرمت دیت تایم می باشدایجاد شده و سپس ستون اولیه از دیتا ست حذف شده است:

---

[۳۳] d
[۳۴] Nan value
[۳۵] DTYYYYMMDD
[۳۶] string

DTYYYYMMDD

```
In [14]: data["Date"]=pd.to_datetime(data["DTYYYYMMDD"].astype(str),format='%Y%m%d')

In [15]: data.Date.agg(["min","max"])

Out[15]: min   2005-06-01
         max   2020-05-31
         Name: Date, dtype: datetime64[ns]

In [16]: data.head()
```

Out[16]:

	Ticker	Per	DTYYYYMMDD	TIME	Open	High	Low	Close	Vol	Openint	Date
0	ABFAR_Share	d	20060227	0	867.65	867.65	867.65	867.65	100	1000	2006-02-27
1	ABFAR_Share	d	20060228	0	867.65	885.00	850.29	864.18	5710	996	2006-02-28
2	ABFAR_Share	d	20060430	0	847.69	847.69	847.69	847.69	13500	977	2006-04-30
3	ABFAR_Share	d	20060501	0	831.21	831.21	831.21	831.21	13500	958	2006-05-01

```
In [17]: data.drop("DTYYYYMMDD",axis=1,inplace=True)
```

برای تجزیه و تحلیل بهتر ورسم نمودار های با مفهوم تر ۱۰ سهام از سهام ها موجود را با استفاده از کد زیر به صورت تصادفی انتخاب میکنم:

Ten stocks are randomly selected and tested

```
In [24]: np.random.seed(101)
         random_num=list(np.random.randint(0,710,size=10))
         random_tricker=[list_of_ticker[x] for x in random_num]

In [25]: for stocke in random_tricker:
             num_of_stock=len(data[data.Ticker==stocke])
             print("We have %d  rows of %s data" %(num_of_stock,stocke))

         We have 2750  rows of SHGOL_Share data
         We have 2234  rows of KHCHARKESH_Share data
         We have 579  rows of TEJARAT_Share data
         We have 786  rows of VAATOS_Share data
         We have 2572  rows of CESHOMAL_Share data
         We have 2881  rows of MADARAN_Share data
         We have 1315  rows of SHSPA_Share data
         We have 277  rows of VAMOALEM_Share data
         We have 579  rows of TEJARAT_Share data
         We have 1942  rows of KEPARS_Share data
```

کمترین ،بیشترین و میانگین مقادیر ستون اوپن اینت را با قطعه کد زیر محاسبه کرد و سطر های دارای این مقادیر را نمایش میدهیم:

Openint

```
In [35]: data.Openint.agg({'min',"max","mean"})
```

```
Out[35]: mean    6.946608e+03
         max     1.271485e+06
         min     1.000000e+00
         Name: Openint, dtype: float64
```

Find the stocks that have the lowest and highest openint values

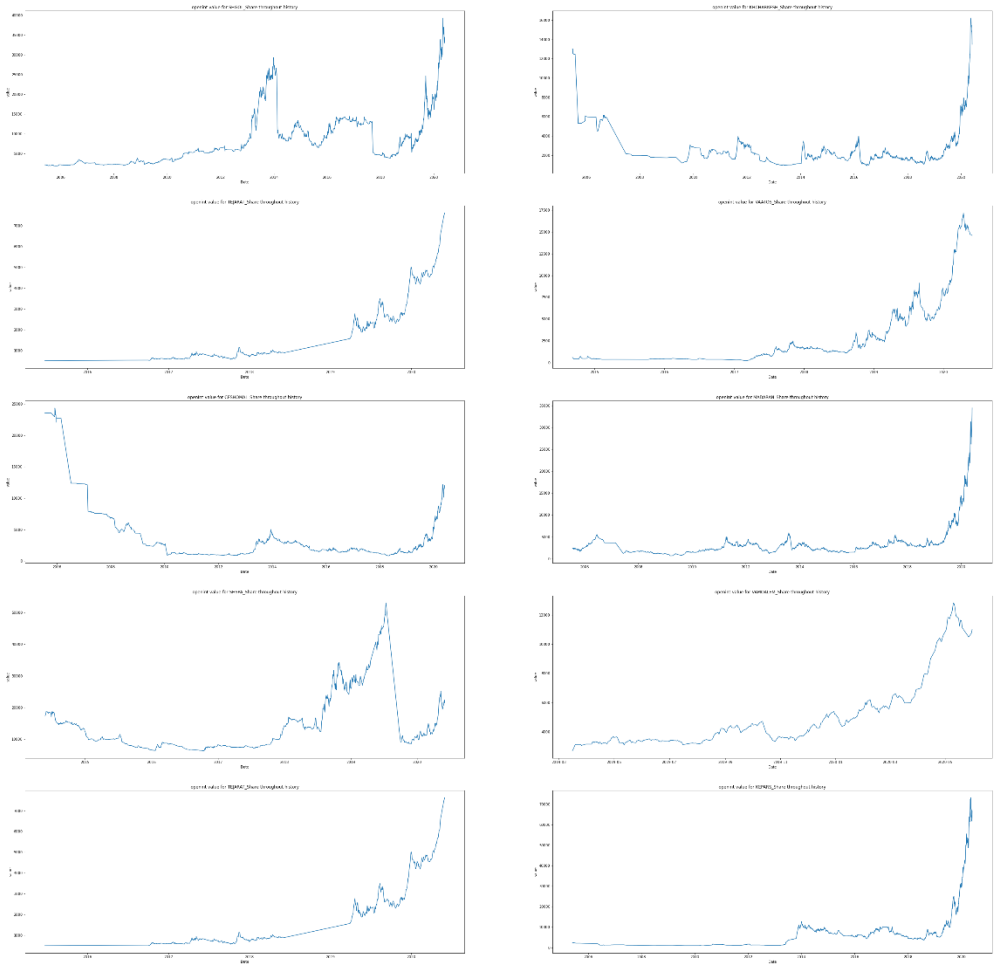
```
In [36]: data[(data.Openint==data.Openint.max()) | (data.Openint==data.Openint.min())]
```

Out[36]:

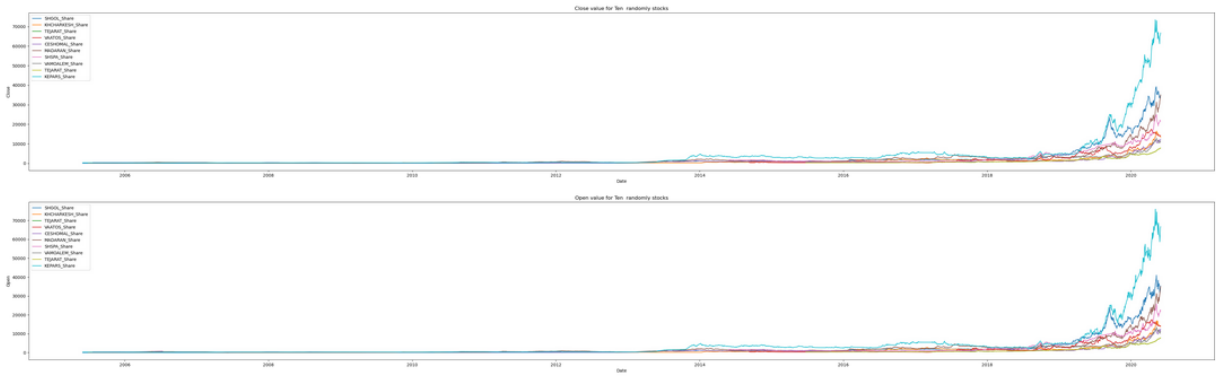
	Ticker	Open	High	Low	Close	Vol	Openint	Date
616574	NEGERBA_Share	1.00	1.00	1.0	1.00	3021	1	2008-01-05
618235	NEREY_Share	1.00	1.00	1.0	1.00	5460000	1	2008-02-26
746030	SHETRAN_Share	4322.26	4322.26	4116.3	4131.41	70458	1271485	2013-10-19

نمودار مقادیر ستون هی کلوز ،اوپن و اوپن اینت بر روی محور زمان را برای سهام های که به صورت رندم انتخاب کردیم با قطعه کدهای زیر رسم میکنیم:

```
In [40]: plt.figure(figsize=(50,50))
         i=1
         for stocke in random_tricker:
             plt.subplot(5,2,i)
             plt.plot(data[data.Ticker==stocke].Date,data[data.Ticker==stocke].Openint.values)
             plt.xlabel('Date')
             plt.ylabel("value")
             plt.title("openint value for %s throughout history"%(stocke))
             i+=1
```



In [46]:
```
plt.figure(figsize=(50,15),dpi=100)
plt.subplot(211)
for stocke in random_tricker:
    plt.plot(data[data.Ticker==stocke].Date,data[data.Ticker==stocke].Close.values)
plt.legend(random_tricker)
plt.xlabel("Date")
plt.ylabel("Close")
plt.title("Close value for Ten  randomly stocks");

plt.subplot(212)
for stocke in random_tricker:
    plt.plot(data[data.Ticker==stocke].Date,data[data.Ticker==stocke].Open.values)
plt.legend(random_tricker)
plt.xlabel("Date")
plt.ylabel("Open")
plt.title("Open value for Ten  randomly stocks");
```



نمی توان از تاریخ و نام سهام ها بعنوان ویژگی های برای اموزش مدل استفاده کرد همچنین نمی شود که این ستون ها را به طور کلی حذف کنیم به همین سبب این ستون ها را با کد زیر به صورت سلسله مراتبی بعنوان ایندکس های دیتاست در نظر میگیریم:

Hierarchical index

```
In [50]: index=list(zip(data.Ticker,data.Date))
         data.index=pd.MultiIndex.from_tuples(index,names=("Ticker","Date"))
         data.drop(["Date","Ticker"],axis=1, inplace=True)

In [51]: data.head()
```

Out[51]:

		Open	High	Low	Close	Vol	Openint
Ticker	Date						
ABFAR_Share	2006-02-27	867.65	867.65	867.65	867.65	100	1000
	2006-02-28	867.65	885.00	850.29	864.18	5710	996
	2006-04-30	847.69	847.69	847.69	847.69	13500	977
	2006-05-01	831.21	831.21	831.21	831.21	13500	958
	2006-05-02	814.72	814.72	814.72	814.72	13500	939

از آنجا که در این پروژه از دو معیار اینرسی[۳۷] و نمره سیلوئیت[۳۸] برای ارزیابی دقت خوشه بندی با مقادیر مختلف k استفاده میکنیم تا بهترین k ممکن را برای خوشه بندی داده ها انتخاب و مورد استفاده قرار دهیم.

[۳۷] inertia
[۳۸] Silhouette score

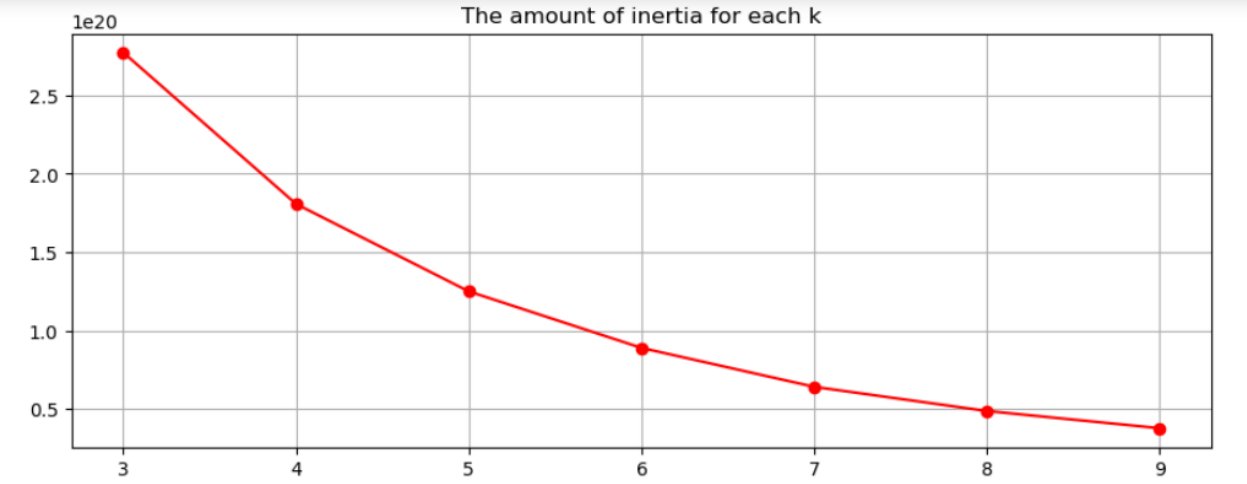
ما الگوریتم کی مینز را با تعدا خوشه های مختلف(در محدوده ۳تا۱۰) اموزش می دهیم و مقدار اینرسی و نمره سیلوئیت را برای هر تعداد خوشه با استفاده از کد زیر محاسبه کرده و روی نمودار ها به نمایش گذاشته ایم:

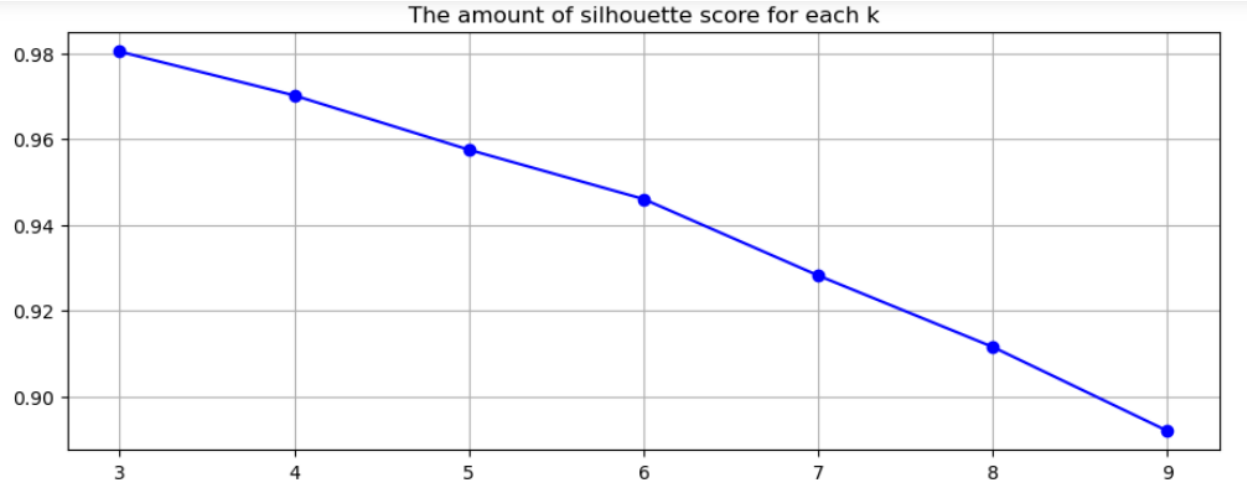
Create Clustering

In [52]:
```
inertia_list=list()
score_list=list()

for k in range(3,10):
    kmn=KMeans(n_clusters=k)
    kmn.fit(data)
    label=kmn.labels_
    inertia_list.append(kmn.inertia_)
    score=silhouette_score(data,label,metric='euclidean',sample_size=5000)
    score_list.append(score)
```

In [53]:
```
plt.figure(figsize=(11,9),dpi=100)
plt.subplot(211)
plt.plot(range(3,10),inertia_list,"ro-")
plt.title("The amount of inertia for each k")
plt.grid()
plt.subplot(212)
plt.plot(range(3,10),score_list,"bo-")
plt.title("The amount of silhouette score for each k")

plt.grid()
```

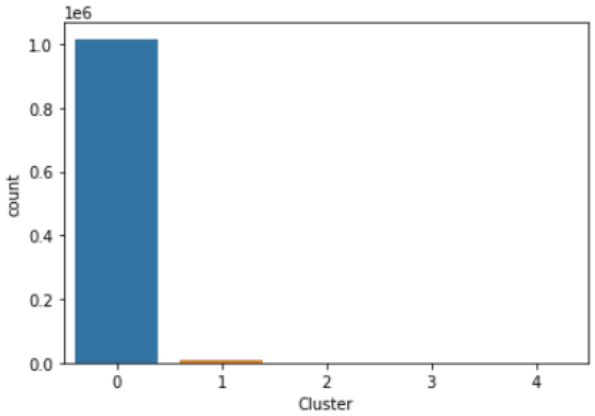
تعداد خوشه ای که مقدار اینرسی را به طور قابل ملاحضه ای کاهش دهد بهترین تعداد خوشه برای خوشه بندی دیتا می باشد با توجه به نمودارهای رسم شده برای دیتا ما بهترین تعداد خوشه براساس معیار اینرسی و نمره سیلوئیت۵ می باشد، با توجه به این موارد ما مدل را با تعداد خوشه ۵ با دیتا ها اموزش می دهیم.



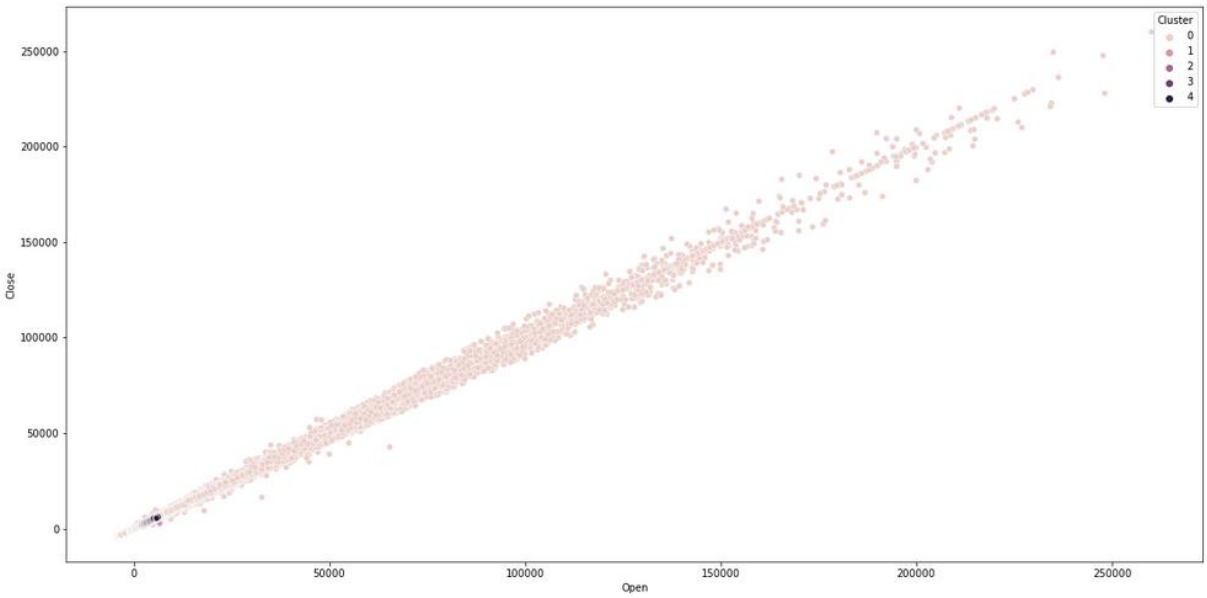


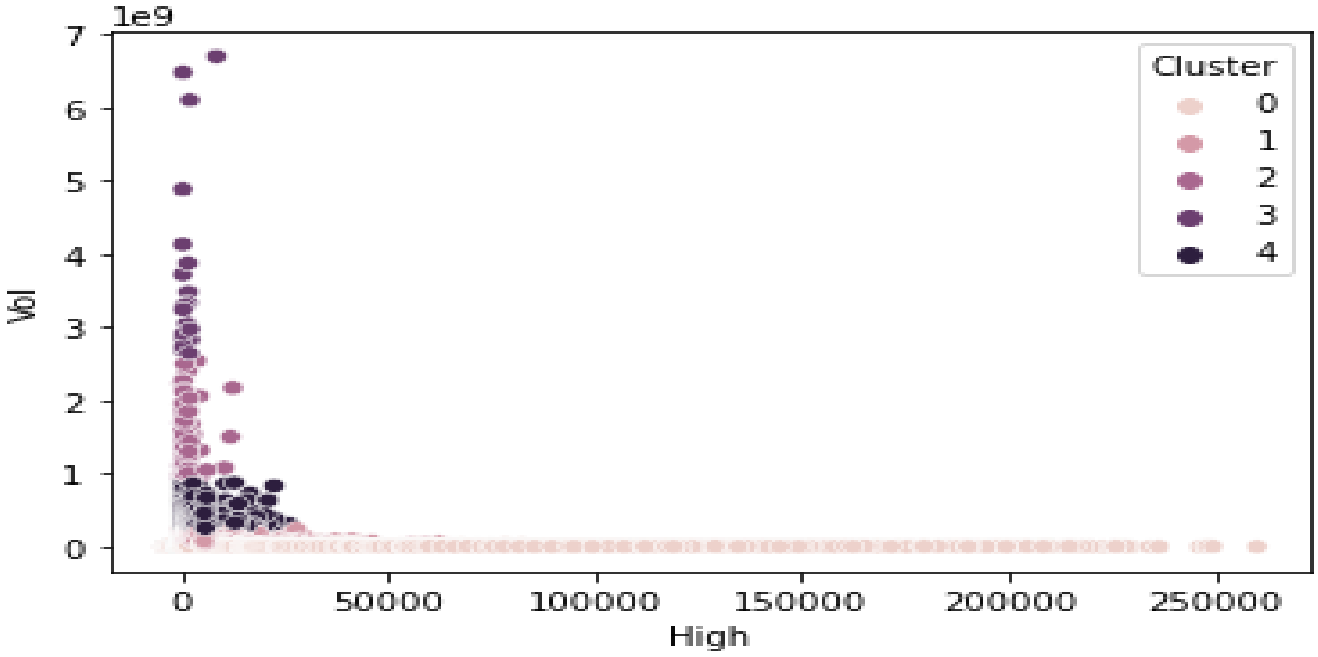
لیبل های که توسط الگوریتم کی مینز به هر داده نسبت داده شده است را بعنوان ستونی به اسم کلاستر به دیتاست اضافه میکنیم وبا کد زیر تعداد داده های موجود درهر خوشه را نمایش می دهیم:

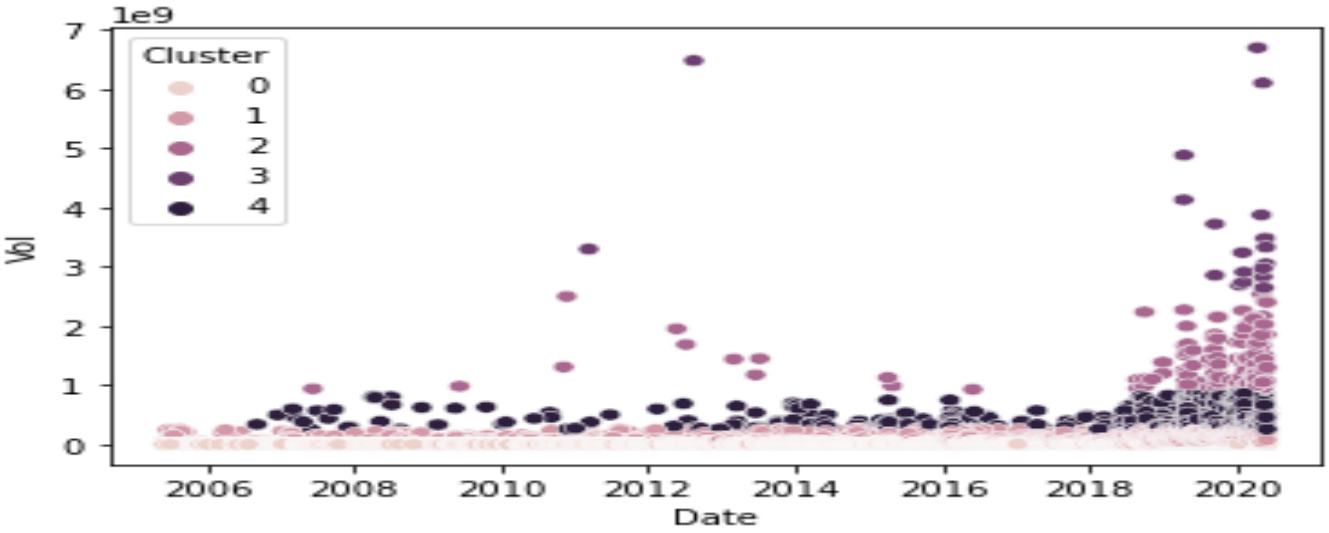
```
In [56]: data.Cluster.value_counts()

Out[56]: 0    1016718
         1      10341
         4        964
         2        134
         3         19
         Name: Cluster, dtype: int64
```



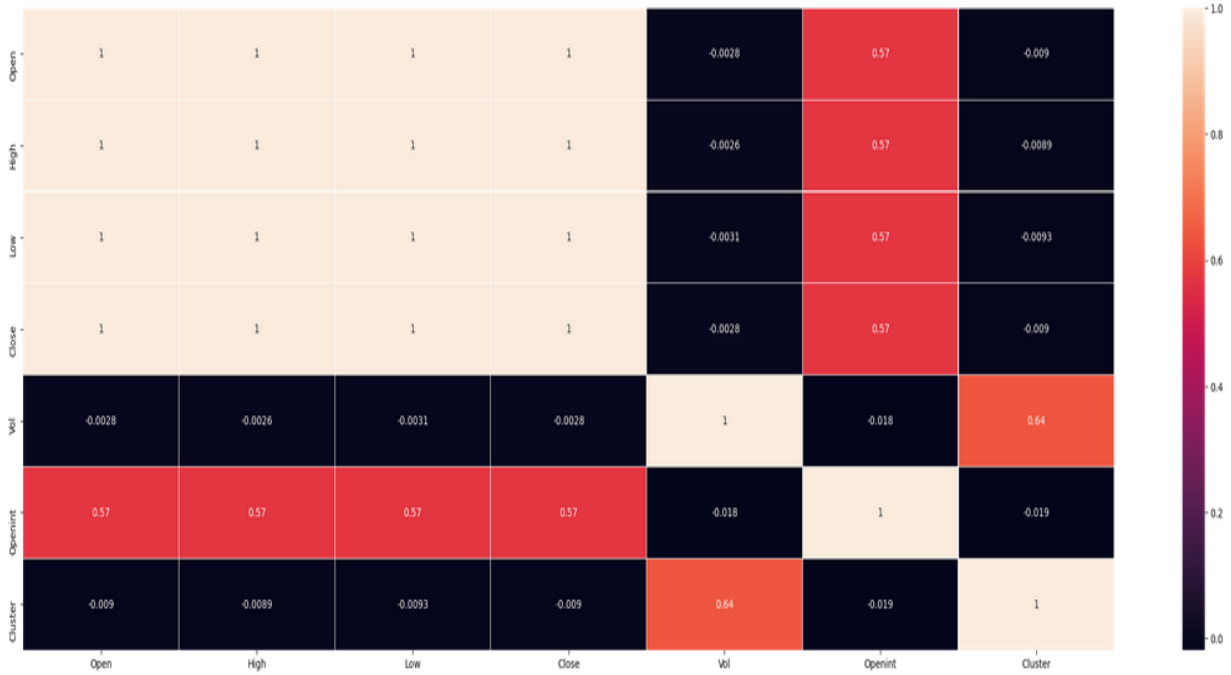
همانطور که مشاهد میکنیم بیشتر داده ها در خوشه شماره صفر هستند.در نمودارهای زیر به خوبی نشان میدهند که دیتا های هر خوشه چه وضعیتی نسبت به سایر خوشه ها دارند و پراکندگی داده های هر خوشه را به نمایش میگذارد :







در نمودار زیر ارتباط هر ویژگی با سایر ویژگی ها و همچنین میزان وابستگی لیبل به هر ویژگی به نمایش گذاشته شده است:

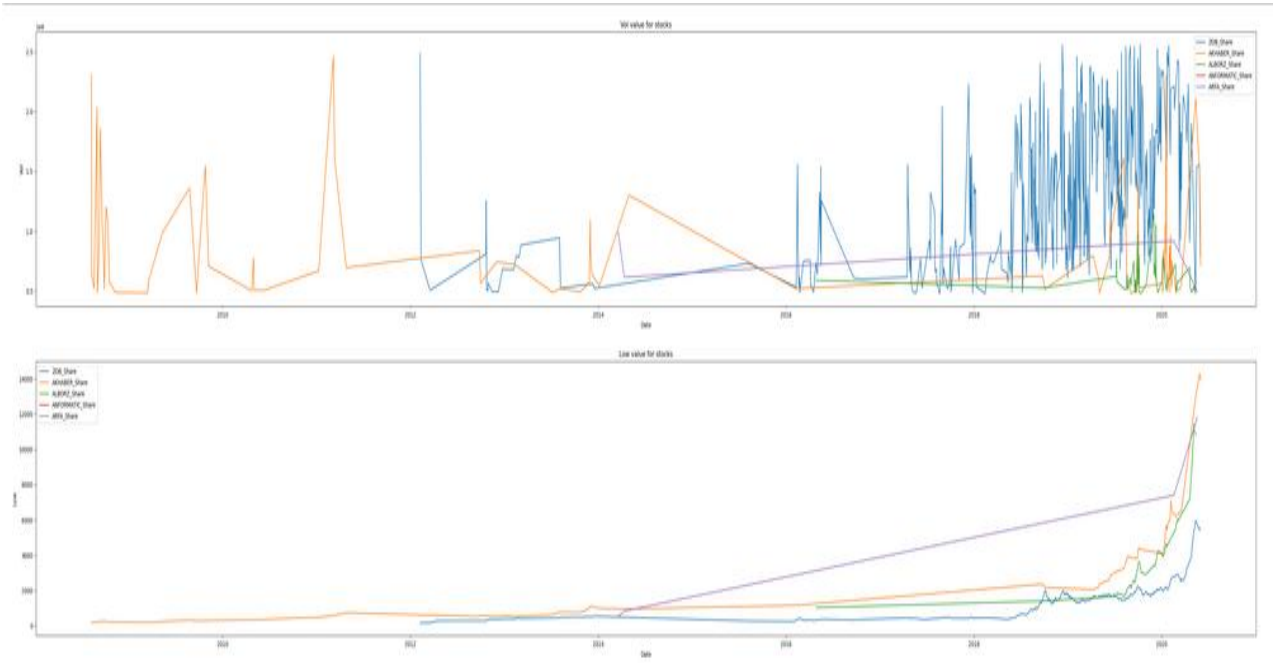


با توجه به نمودار بالا ستون حجم بیشترین تاثیر را در انتخاب خوشه دارد.

ما با استفاده از تکنیک های خوشه بندی توانستیم سهام های که رفتار مشابه بهم دارند را شناسایی کرده و در یک کلاس قرار دهیم.ما می خواهیم ببینیم ایا با داشتن مقادیر مختلف سهام های موجود در یک کلاستر می توان همین مقادیر را برای سهام های دیگر در این کلاستر پیش بینی کرد برای این منظور باید از الگوریتم های رگرسیون استفاده کنیم.

ساخت مدل رگرسیون روی داده های موجود در خوشه شماره یک:

ما مدل های رگرسیون مختلفی را روی داده خوشه شمار یک آموزش داده و سپس تست کردیم خوشه شماره یک شامل ۲۷۶ سهام می باشد نمودار زیر نشان داهند مقادیر موجود در ستون های حجم معاملات و کمترین قیمت معاملات بر روی محور زمان برای چند سهام موجود در کلاستر یک می باشد:



در اینجا متغییر هدف ما مقدار بسته شدن سهام می باشد در نتیجه با استفاده از کد زیر داده ها را به دو قسمت X,Yتقسیم می کنیم، سپس داده های آموزش و تست را از هم تفکیک کرده ایم، داده های تست ما قیمت هاس۵ سهام می باشد که به صورت تصادفی انتخاب شده اند و داده های آموزشی ما ۲۷۱ سهام دیگر که در خوشه شماره یک وجود دارند می باشند:

train test split

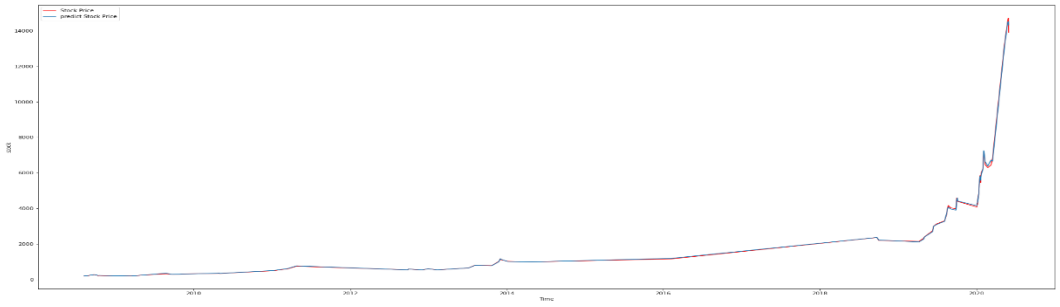
```
In [77]: x=cluster_one.drop("Close",axis=1)
         y=cluster_one["Close"]
```

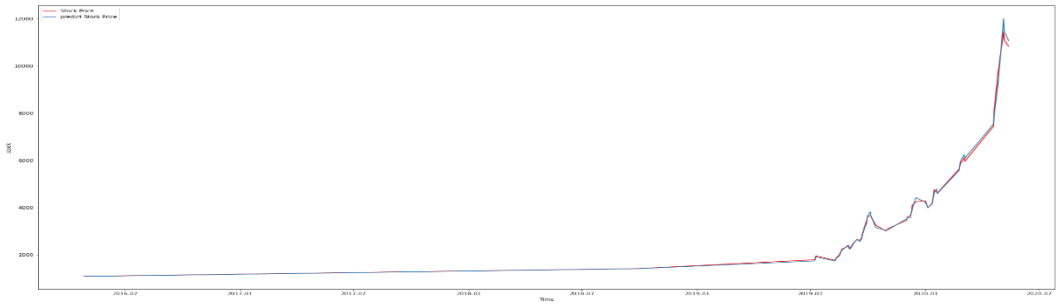
```
In [106]: xtrain=x.drop(["ZOB_Share","AKHABER_Share", 'ALBORZ_Share', 'ANFORMATIC_Share', 'ARFA_Share'],axis=0)
          xtest=x[(x.index=="ZOB_Share")| (x.index=="AKHABER_Share")| (x.index=="ALBORZ_Share")| (x.index=="ANFORMATIC_Share")|(x.index=

          ytrain=y.drop(["ZOB_Share","AKHABER_Share", 'ALBORZ_Share', 'ANFORMATIC_Share', 'ARFA_Share'],axis=0)
          ytest=y[(y.index=="ZOB_Share")| (y.index=="AKHABER_Share")| (y.index=="ALBORZ_Share")| (y.index=="ANFORMATIC_Share")|(y.index=
```

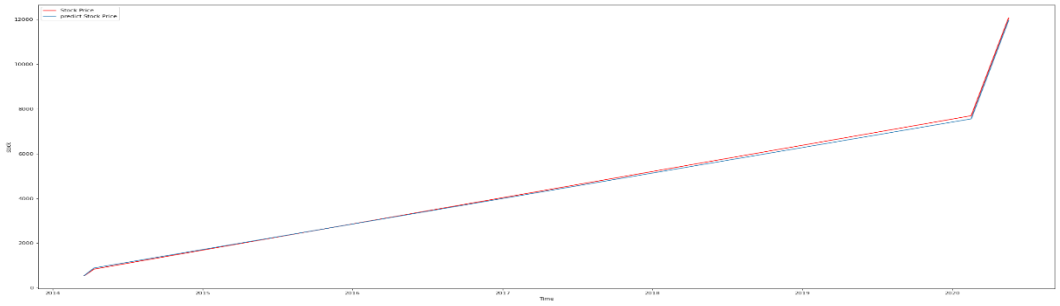
چهار الگوریتم رگرسیون خطی ریدگ رگرسیون و لاسو رگرسیون و درخت تصمیم رو داده ها اموزش داده شده و دقت آنها سنجیده شده است که در زیر نتایج حاصل از هر الگوریتم به نمایش گذاشته شده است:

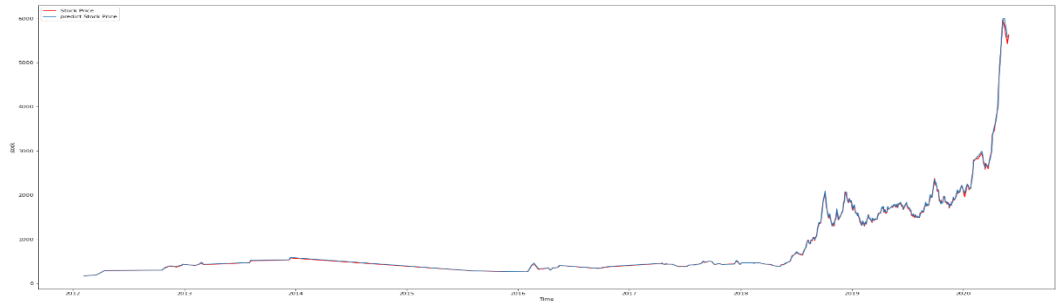
رگرسیون خطی:با دقت های ۹۹,۹۰روی داده های ترین و ۹۹,۹۲ رو داده های تست و میزان خطای ۳۳۳۸,۳٤٦ با معیار حداقل مربعات خطا[۳۹] باشد.نمودار زیر نشان دهنده ای اختلاف مقادیر واقعی ومقادیر پیش بینی شده توسط مدل رگرسیون می باشد:

[۳۹] Mean squared error


Stock Price
predict Stock Price
Time

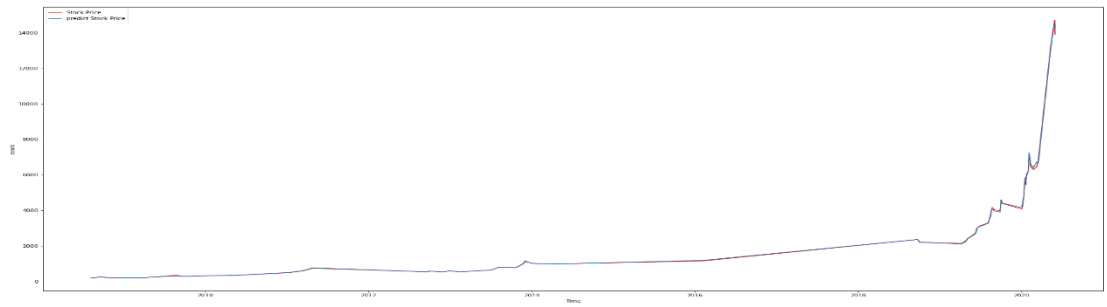

Stock Price
predict Stock Price
Time

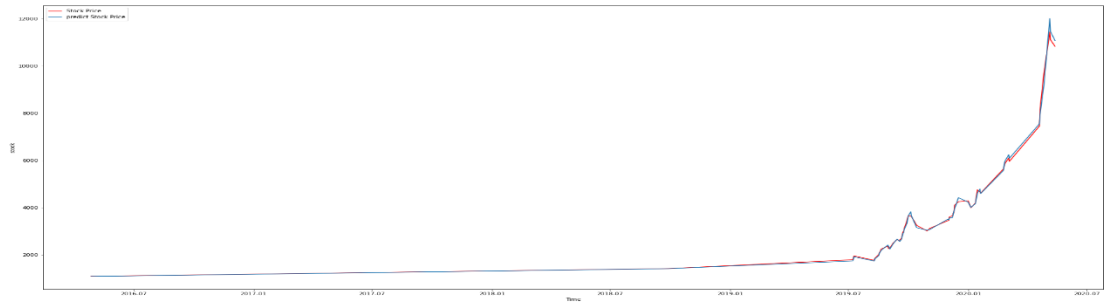

Stock Price
predict Stock Price
Time

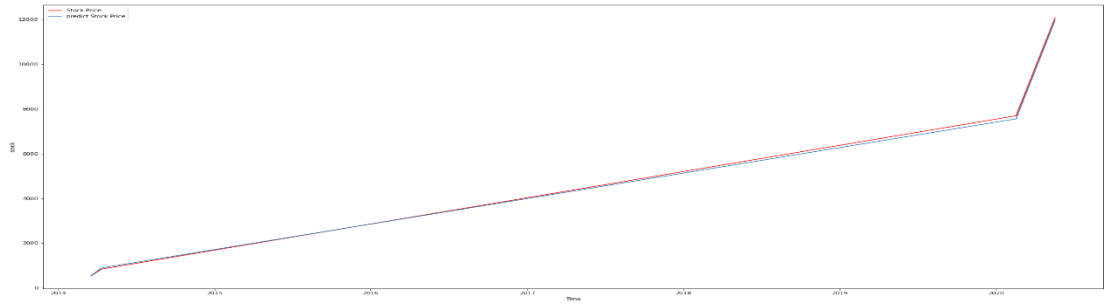

Stock Price
predict Stock Price
Time

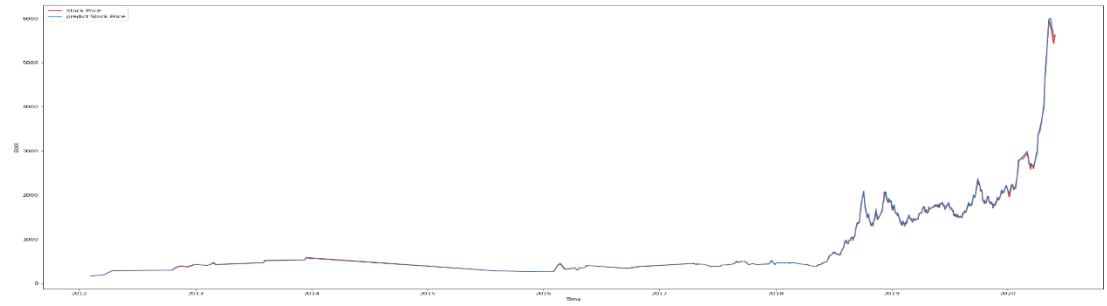
ریدگ:[4]با دقت های ۹۹,۹۰روی داده های ترین و ۹۹,۹۲ رو داده های تست و میزان خطای ۳۳۳۸,۳٤٦ با معیار حداقل مربعات خطا می باشد.نمودارهای زیر نشان دهنده ای اختلاف مقادیر واقعی ومقادیر پیش بینی شده توسط مدل ریدگ می باشد:

[4]Ridge

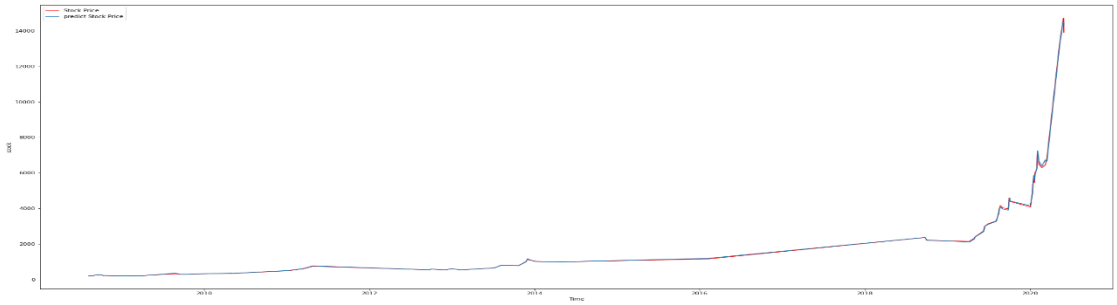

Stock Price
predict Stock Price
Time

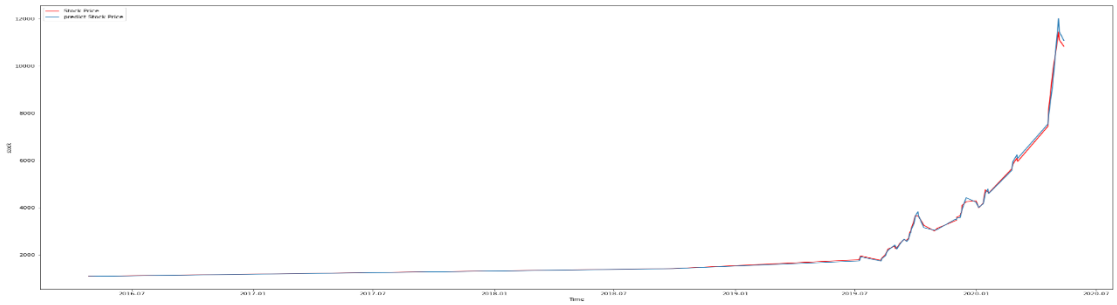

Stock Price
predict Stock Price
Time

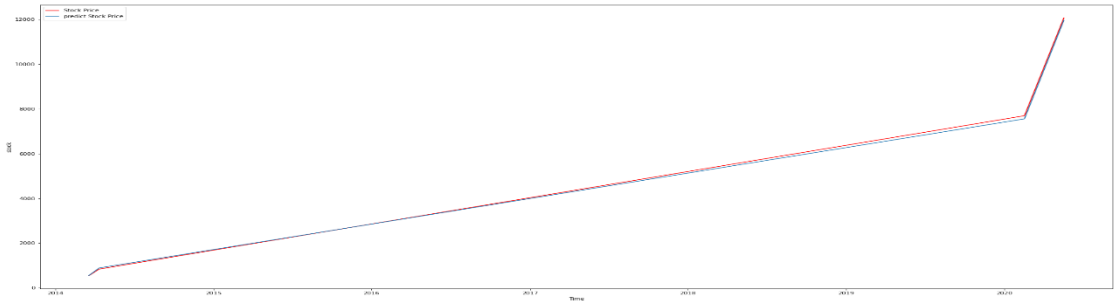

Stock Price
predict Stock Price
Time

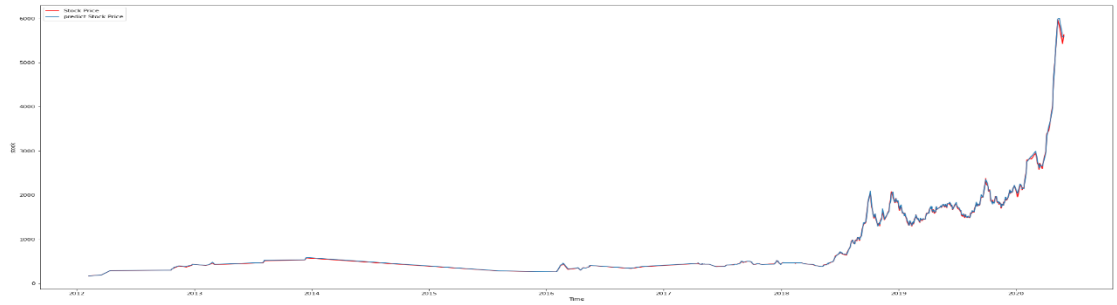

Stock Price
predict Stock Price
Time

لاسوو: با دقت های ۹۹,۹۰روی داده های ترین و ۹۹,۹۲ رو داده های تست و میزان خطای ۳۳۳۸,۳٤٦ با معیار حداقل مربعات خطا می باشد.نمودارهای زیر نشان دهنده ای اختلاف مقادیر واقعی ومقادیر پیش بینی شده توسط مدل لاسو می باشد:

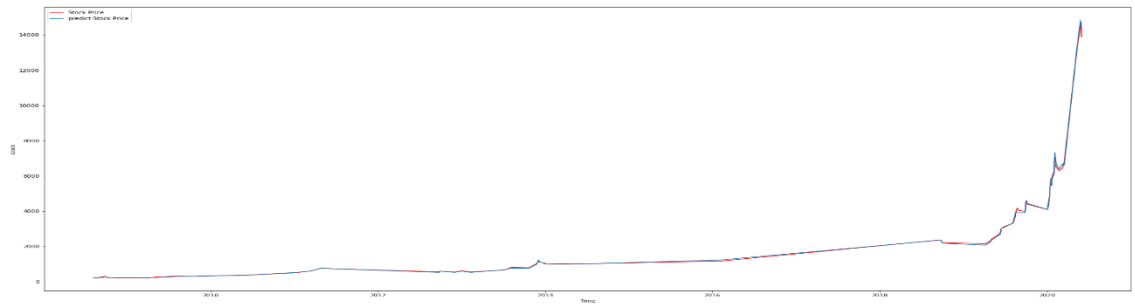

Stock Price
predict Stock Price
Time

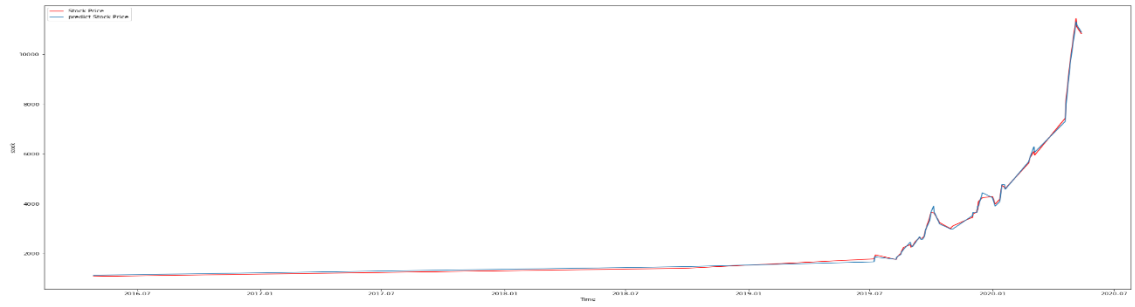

Stock Price
predict Stock Price
Time

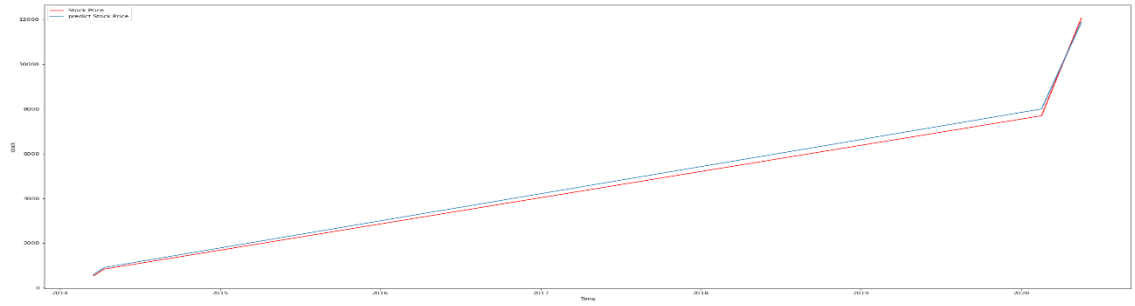

Stock Price
predict Stock Price
Time

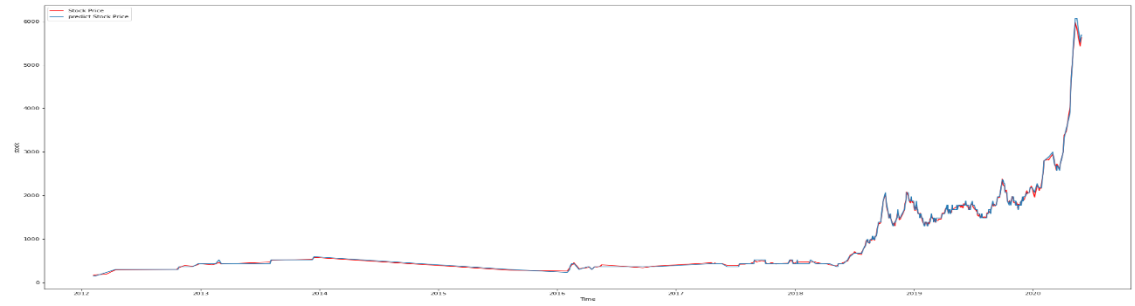

Stock Price
predict Stock Price
Time

درخت تصمیم: با دقت های۹۹,۹٤روی داده های ترین و دقت ۹۹,۸۹ رو داده های تست و میزان خطای ۳۸۸۸,۸۳۸ با معیار حداقل مربعات خطا می باشد.نمودارهای زیر نشان دهنده ای اختلاف مقادیر واقعی ومقادیر پیش بینی شده توسط مدل درخت تصمیم می باشد:


Stock Price
predict Stock Price
Time


Stock Price
predict Stock Price
Time


Stock Price
predict Stock Price
Time


Stock Price
predict Stock Price
Time

خلاصه نتایج:

ما با استفاده از الگوریتم های خوشه بندی کی مینز داده ها را به ۵ خوشه خوشه بندی کردیم که این الگوریتم با این تعداد خوشه عملکرد خوبی در خوشه بندی داده ها داشته سپس الگوریتم های رگرسیون ریدگ لاسو و درخت تصمیم را با استفاده از داده های موجود در یکی از خوشه ها به این منظور که دقت الگوریتم خوشه بندی را در شناسایی روند رفتار سهام ها بسنجیم آموزش دادیم و نتایج حاصل از آموزش الگوریتم های رگرسیون بدین شرح بوده است :الگوریتم درخت تصمیم با دقت های۹۹,۹٤روی داده های ترین و دقت ۹۹,۸۹ رو داده های تست و میزان خطای ۳۸۸۸,۸۳۸ ، لاسوو با دقت های ۹۹,۹۰روی داده های ترین و ۹۹,۹۲ رو داده های تست و میزان خطای ۳۳۳۸,۳٤٦ ،ریدگ با دقت های ۹۹,۹۰روی داده های ترین و ۹۹,۹۲ رو داده های تست و میزان خطای ۳۳۳۸,۳٤٦.معیاری که ما در این پژوهش برای محاسبه میزان خطای الگوریتم های رگرسون استفاده کردیم معیار حداقل مربعات خطا[۱٤]می باشد.همان ظور که گفته شده نتیجه آموزش الگوریتم های رگرسیون مدل های با دقت بالا و میزان خطای زیاد بوده است که این مسئله نشان میدهد که الگوریتم خوشه بندی به خوبی عملکرده است ومیزان بالای خطا ممکن است ناشی از عدم بهینه سازی پارامتر های الگوریتم های رگرسیون باشد.

[۱٤] Mean squared error

پیشنهادات آتی:

برای استفاده بهتر از هوش مصنوعی در معاملات بهتر است از ترکیب چندین سیستم میتنی بر هوش مصنوعی و یک تیم از متخصصین این حوزه با هم همکاری داشته باشند زیر اثبات شده است که ترکیب هوش مصنوعی و انسان همیشه نتایج خارق العاده ای را به همراه داشته است.

سیستم های پیشنهادی هوش مصنوعی که میتوانددر معاملات کمک بسزایی را به تحلیل گران بکنند به شرح زیر می باشند:

۱- از آنجا اخبار مربوط به اقتصاد، سیاست و ... ،و نقطه نظرات اشخاص سر شناس می تواند روی روند بازار اثرا داشه باشد باید سیستم های مبتنی پردازش زبان طبیعی توسعه داده شود تا اخبار را رسد کرد و تاثیر آنها را در اختیار متخصصین و سایر الگوریتم ها قرار دهد.
۲- سیتمی برای تحلیل رفتار نهنگ های بازار های مالی از آنجای که نهنگ های بازار مالی با رفتار های خود سعی میکنند روند بازار را به سمت مورد نظر خود هدایت کنند نباید از تاثیر رفتار های آنها غافل بود.
۳- سیستمی برای تحلیل رفتار سهام های بورس و ارتباط آنها با یکدیگر از الگوریتم های خوشه بندی برای توشعه این سیستم استفاده شود

٤- سیستم های مجزای برای پیش بینی دقیق قیمت سهام ها با استفاده از الگوریتم های رگرسون بخصوص الگوریتم شبکه های عصبی بازگشتی و استفاده ازالگوریتم ازدحام ذرات برای بهینه سازی پارامتر های شبکه عصبی بازگشتی

هوش مصنوعی درآینده نزدیک تمام جنبه های زندگی بشر را متحول می سازد،از آنجا که حوزه معاملات اقتصادی و بازار بورس یکی از حوزه های پیچیده می باشد که نیازمند تحلیل های دقیق ،هوشمندانه و منطقی است به همین سبب استفاده از هوش مصنوعی میتواند درهای جدیدی به روی دنیای اقتصاد باز کند .با وجود چنین سیستم های ریسک معاملات کاهش پیدا می کند،درنتیجه معامله گر با اطمینان بهتری معاملات خود را انجام دهد.هوش مصنوعی در زمینه معاملات اقتصادی و بازار بورس میتواند این کاربرد ها را داشته باشد:

۱-کمک به داشتن فرآیندی اتوماتیک و دقیق تر برای معاملات

۲-پیش بینی روند پیش رو به کمک تجزیه و تحلیل داده های گذشته

۳-با نظارت بر بازار و تجریه و تحلیل آن این سیستم های میتوانن فرصت های معاملاتی برای داشتن خرید و فروش های کارآمد برای معامله گران ایجاد کنند.