

دانشگاه صنعتی شریف

دانشکده مهندسی صنایع

پروژه درس داده‌کاوی

**تعیین رنگ شمع در روز آینده از طریق تحلیل پارامترهای بازار مالی فارکس به کمک تکنیک­های داده­کاوی**

نگارش:

آرش سرایی 96206852

نیلوفر مبصّری 97208855

استاد درس:

جناب آقای دکتر مجید خدمتی

نیمسال اول 99-98



فهرست

[فصل اول: مقدمه 5](#_Toc31027615)

[فصل دوم: مروری بر ادبیات 7](#_Toc31027616)

[2-1 مقدمه 8](#_Toc31027617)

[2-2 اصول و تعاریف داده‌کاوی 8](#_Toc31027618)

[**2-2-1 تاريخچه داده‌كاوي** 8](#_Toc31027619)

[**2-2-2 فرآيند داده‌كاوي** 9](#_Toc31027620)

[**2-2-3 كاربردهاي داده‌كاوي** 9](#_Toc31027621)

[**2-2-4 يادگيري ماشين** 10](#_Toc31027622)

[2-3 روش‌های داده‌کاوی 11](#_Toc31027623)

[2-4 تحلیل داده‌های مالی ازدیدگاه‌های گوناگون 12](#_Toc31027624)

[**2-4-1 سير زماني تحقيق** 12](#_Toc31027625)

[**2-4-2 مقالات انجام شده** 13](#_Toc31027626)

[فصل سوم: پیش‌پردازش داده‌ها 15](#_Toc31027627)

[3-1 مقدمه 16](#_Toc31027628)

[3-2 پیش پردازش داده‌ها 16](#_Toc31027629)

[فصل چهارم: شناسایی ویژگی‌های مورد نظر جهت داده‌کاوی 17](#_Toc31027630)

[4-1 مقدمه 18](#_Toc31027631)

[4-2 معرفی ویژگی‌های موجود 18](#_Toc31027632)

[فصل پنجم: انتخاب ویژگی‌های مطلوب 21](#_Toc31027633)

[5-1 متدولوژي‌هاي موجود جهت انتخاب ويژگي 22](#_Toc31027634)

[5-2 انتخاب ويژگي‌هاي با بيشترين اثرگذاري 23](#_Toc31027635)

[فصل ششم: داده‌کاوی و شناسایی الگوهای پنهان در داده‌ها، و ارائه نهایی مدل‌هاي طبقه‌بندي 26](#_Toc31027636)

[فصل هفتم: نتیجه‌گیری 32](#_Toc31027637)

[پیوست: کد نرم‌افزار و توضیح درباره نرم‌افزار مورد استفاده 34](#_Toc31027638)

[ نرم‌افزار مورد استفاده جهت پياده سازي مدل 34](#_Toc31027639)

[ كد نرم‌افزار 34](#_Toc31027640)

[منابع و مراجع 35](#_Toc31027641)

فهرست اشكال و جداول

[شكل2-1 گام‌هاي فرآيند داده‌كاوي 9](#_Toc31027589)

[شکل 2-2 کلاس‌های مختلف یادگیری ماشین 11](#_Toc31027590)

[شكل 2-3 دسته‌بندي روش‌هاي مختلف داده‌كاوي 12](#_Toc31027591)

[جدول 4-1 ويژگي‌هاي موجود در مجموعه داده‌ي مورد بررسي پژوهش 18](#_Toc31027592)

[شكل 4-1 اطلاعات ويژگي‌هاي مورد استفاده در پژوهش 20](#_Toc31027593)

[شكل 5-1 روش‌هاي انتخاب ويژگي 22](#_Toc31027594)

[شكل 5-2 مزايا و معايب روش‌هاي انتخاب ويژگي 23](#_Toc31027595)

[جدول 5-1 مرتب‌سازي ويژگي‌هاي موجود در مسئله بر اساس ميزان اهميت 23](#_Toc31027596)

[شكل 7-1 پياده‌سازي مراحل تشريح شده در نرم افزار پايتون 28](#_Toc31027597)

[شكل 7-2 نتايج حاصل از مدل 29](#_Toc31027598)

[شكل 7-3 ماتريس درهم‌ريختگي 30](#_Toc31027599)

# فصل اول: مقدمه

با پیشرفت سریع فناوری اطلاعات[[1]](#footnote-1)، بشر شاهد یک رشد انفجاری در تولید داده و ظرفیت‌های گردآوری و ذخیره‌سازی آن در دامنه‌های گوناگون بوده است. در جهان کسب‌و‌کار، پایگاه‌داده‌های[[2]](#footnote-2) بسیار بزرگی برای تراکنش‌های تجاری وجود دارند که توسط خرده‌فروشان و یا در تجارت الکترونیک[[3]](#footnote-3) ساخته شده‌اند. از سوی دیگر، همه روزه حجم عظیمی از داده‌های علمی در زمینه‌های گوناگون تولید می‌شوند. در چنین شرایطی، تحلیل بدنه بزرگ داده‌ها به شکل قابل درک و کاربردی، یک مساله چالش برانگیز است.

داده‌کاوی[[4]](#footnote-4) این مساله را با فراهم کردن روش‌ها و نرم‌افزارهایی برای خودکارسازی تحلیل‌ها و اکتشاف مجموعه داده‌های بزرگ و پیچیده حل می‌کند. امروزه استفاده از تکنیک­های این حوزه به شدت در حال افزایش است. این موضوع در کنار یادگیری ماشین سبب شده است تا مسائل پیچیده به سادگی تحلیل گشته و اشتیاق به کشف دانش و الگوهای جدید از داده­های خام، بیش از همیشه مورد توجه واقع گردد.در این پژوهش سعی شده است تا نگاهی متفاوت و کاربردی نو برای این حوزه معرفی گردد. اگرچه پیش از این بر روی مدل­های مشابه کار شده است اما در این پژوهش به یک تحلیل عمیق نسبت به این موضوع پرداخته شده و جوانب مختلف آن پوشش داده شده است.

هدف اصلی دنبال شده در این پژوهش پیدا کردن راهی برای پیش­بینی رنگ شمع روز آینده از طریق تحلیل پارامترهای موجود در مجموعه داده‌های روزانه‌ی جفت ارز دلار-ین[[5]](#footnote-5) در بازار فارکس[[6]](#footnote-6) می‌باشد. در همین راستا این پژوهش در هفت فصل گردآوری شده است. ابتدا به ادبیات موضوع پرداخته شده است تا با فعالیت‌های انجام شده در این زمینه آشنایی اولیه‌ای صورت بگیرد. این موضوع در فصل دوم پیگیری شده است. در قدم بعد، به تشریح داده‌های مورد استفاده در این پژوهش پرداخته شده و متغیرهای به کار رفته هر یک معرفی و نوع آن مشخص شده است. در فصل چهارم به تحلیل اولیه‌ای در خصوص نحوه رفتار هر متغیر به تنهایی و نیز ارتباط هر یک با دیگری پرداخته شده است تا بدین طریق ضمن درک درست از شرایط موجود مدلسازی بهتر و دقیق‌تری صورت بگیرد، همچنین در این فصل فعالیت‌های لازم جهت آماده‌سازی داده‌ها برای تحلیل‌های آتی و استخراج دانش انجام شده است. در فصل پنجم در راستای دستیابی به دقت بالاتر در پیش‌بینی و طبقه‌بندی داده‌های آتی به انتخاب پارامترهای منتخب از میان پارامترهای موجود پرداخته شده است. فصل ششم به عنوان فصل کلیدی این پژوهش به دنبال پاسخ مناسب برای پرسش اصلی این مطالعه یعنی پیدا کردن راهی برای پیش­بینی رنگ شمع روز آینده از طریق تحلیل پارامترهای موجود در مجموعه داده‌های روزانه‌ی جفت ارز دلار-ین[[7]](#footnote-7) در بازار فارکس[[8]](#footnote-8) می‌باشد. بدین منظور روش‌های مختلفی به کار گرفته شده‌اند تا در نهایت بتوان مدل یا مدل‌های مناسبی برای این هدف تعیین نمود. در انتها و در فصل هفتم نیز نتیجه‌گیری این پژوهش ارائه شده است. همچنین کد نرم‌افزار استفاده شده برای این پژوهش و توضیحات تکمیلی در ارتباط با نرم‌افزار مورد استفاده در پیوست آمده است.

# فصل دوم: مروری بر ادبیات

## 2-1 مقدمه

در این فصل بر ادبیات موضوع مورد مطالعه مرور مختصری شده است. در بخش اول به بررسی اصول داده‌کاوی و تعاریف مرتبط با آن پرداخته شده است.در ادامه به روش‌ها و تکنیک‌های قابل استفاده در این مسیر و نیز فرآیند کلی آن اشاره شده است. در بخش پایانی موضوع بازارهای مالی و پارامترهای آن به صورت مختصر مورد مطالعه قرار گرفته است و در نهایت روش‌های به کار گرفته شده در این راستا توسط سایر محققین بررسی شده است.

ما در جهانی زندگی می کنیم که روزانه مقادیر عظیمی از داده ها در آن جمع‌آوري می شوند. تجزیه و تحلیل چنین داده هایی یک نیاز مهم و ضروري است.(ديتا ماينينگ-كانسپتز اند تكنيكز)

## 2-2 اصول و تعاریف داده‌کاوی

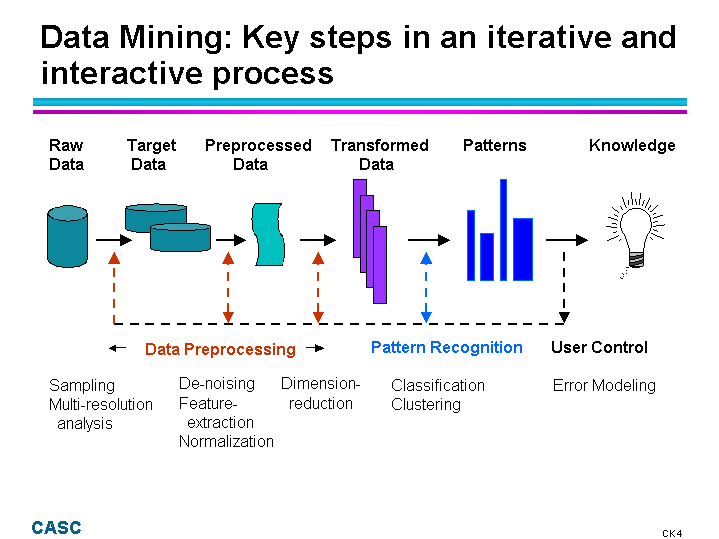
ما در جهانی زندگی می کنیم که روزانه مقادیر عظیمی از داده ها در آن جمع‌آوري می شوند. تجزیه و تحلیل چنین داده هایی یک نیاز مهم و ضروري است(ديتا ماينينگ-كانسپتز اند تكنيكز). داده‌کاوی[[9]](#footnote-9) این مساله را با فراهم کردن روش‌ها و نرم‌افزارهایی برای خودکارسازی تحلیل‌ها و اکتشاف مجموعه داده‌های بزرگ و پیچیده حل می‌کند. پژوهش‌ها در زمینه داده‌کاوی در گستره وسیعی از موضوعات شامل آمار، علوم کامپیوتر، یادگیری ماشین[[10]](#footnote-10)، مدیریت پایگاه داده[[11]](#footnote-11) و بصری‌سازی داده‌ها[[12]](#footnote-12) دنبال می‌شود. روش‌های داده‌کاوی و یادگیری، در زمینه‌هایی غیر از آمار نیز توسعه داده شده‌اند، که از جمله آن‌ها می‌توان به یادگیری ماشین و پردازش سیگنال[[13]](#footnote-13) اشاره کرد.

### **2-2-1 تاريخچه داده‌كاوي**

واژه داده کاوی تا اوایل دهه 90 میلادی مفهومی نداشت و بکار برده نمی شد. در دهه 60 میلادی متخصصان آمار به جای استفاده از کلمه تحلیل داده، از واژه های دیگری مانند صید داده[[14]](#footnote-14) و لایبروبی داده[[15]](#footnote-15) استفاده می کردند. اصل اصطلاح و واژه داده کاوی در ابتدا دهه 90 میلادی مورد استفاده قرار گرفت. در دهه 60 میلادی و پیش از آن زمینه هایی برای ایجاد سیستم های جمع آوری و مدیریت داده ها ایجاد شد و تحقیقاتی در این زمینه انجام پذیرفت که منجر به معرفی و ایجاد سیستم های مدیریت پایگاه داده شد. توسعه سیستم های پایگاهی پیشرفته در دهه 80 و ایجاد پایگاه های شی گرا، کاربردگرا و فعال باعث توسعه همه جانبه و کاربردی شدن این سیستم ها در سراسر جهان گردید. بدین ترتیب سیستم های مدیریت پایگاه داده‌ای همچونDB2 ، Oracle، Sybase و غیره ایجاد شدند و حجم زیادی از داده ها توسط این سیستم ها مورد پردازش قرار گرفت. شاید بتوان مهمترین عامل در معرفی داده کاوی را مبحث کشف دانش از پایگاه داده [[16]](#footnote-16)دانست بطوری که در بسیاری از موارد کشف دانش از پایگاه داده و داده کاوی بصورت مترادف بکار برده می شوند. الگوریتم های داده کاوی در دهه اخیر با سرعت بسیار زیاد در حال توسعه هستند.

### **2-2-2 فرآيند داده‌كاوي**

داده‌کاوی که با عنوان کشف دانش از داده نیز شناخته شده است، به فرایند استخراج اطلاعات و دانش از داده‌های موجود در پایگاه‌داده یا انبارداده اطلاق مي‌شود. فرآیند داده‌کاوی شامل چندین گام است. این فرآیند از داده‌های خام آغاز می‌شود و تا شکل‌دهی دانش جدید ادامه دارد(شكل2-1).



شكل2-1 گام‌هاي فرآيند داده‌كاوي

### **2-2-3 كاربردهاي داده‌كاوي**

داده کاوی تقریبا درتمامی صنایع ، کسب و کار ها و تجارت ها و بخش های روزمره زندگی ما کاربرد دارد. در اصل هر جایی که ما با داده و داده های حجیم سروکار داریم ، داده کاوی می تواند به کمک ما بیاید. به عنوان مثال :

* پیدا کردن بازار هدف برای کسب و کار ها
* کشف الگور های رفتاری خرید مشتری در فروشگاه ها و کسب و کار ها
* تحلیل سبد خرید
* شناسایی مشتریان وفادار
* آنالیر دقیق نیاز های مشتریان
* پیش بینی فروش
* دسته بندی مشتریان بر اساس ملیت ، نژاد ، زبان ، موقعیت مکانی و …
* در زمینه های مختلف بانک داری مانند پیش بینی الگو های کلاه برداری در بانکداری
* مدیریت ریسک
* علم پزشکی ( در بخش های مختلف مثل پیشگیری سرطان ، تشخیص بیماران ، درمان بیماران ، تاثیر اثر دارو بر بیمار و …)
* علم اقتصاد ( در بخش های مختلف مانند پیش بینی آینده ، مدیریت سرمایه و… )
* علم ژنتیک
* شناسایی مجرمان

### **2-2-4 يادگيري ماشين**

یادگیری ماشین يكي از موضوعات پركاربرد در زمينه داده‌كاوي است كه زیر مجموعه‌ای از هوش مصنوعی به حساب مي‌آيد. با استفاده از تکنیک‌های یادگیری ماشین، کامپیوتر، الگوهای موجود در داده‌ها (اطلاعات پردازش شده) را یادگرفته و می‌تواند از آن استفاده کند.

یادگیری ماشین خود به سه کلاس تقسیم می شود: یادگیری نظارت‌یافته[[17]](#footnote-17)، بدون نظارت[[18]](#footnote-18) و تقویتی[[19]](#footnote-19). در بندهای زیر، این سه روش یادگیری با روش‌های رایج در هر کلاس مورد بحث قرار می‌گیرد و همچنین در شکل 1-1 نگاهی کلی به این 3 کلاس شده است.

#### **2-2-4-1 یادگیری ماشین نظارت‌یافته**

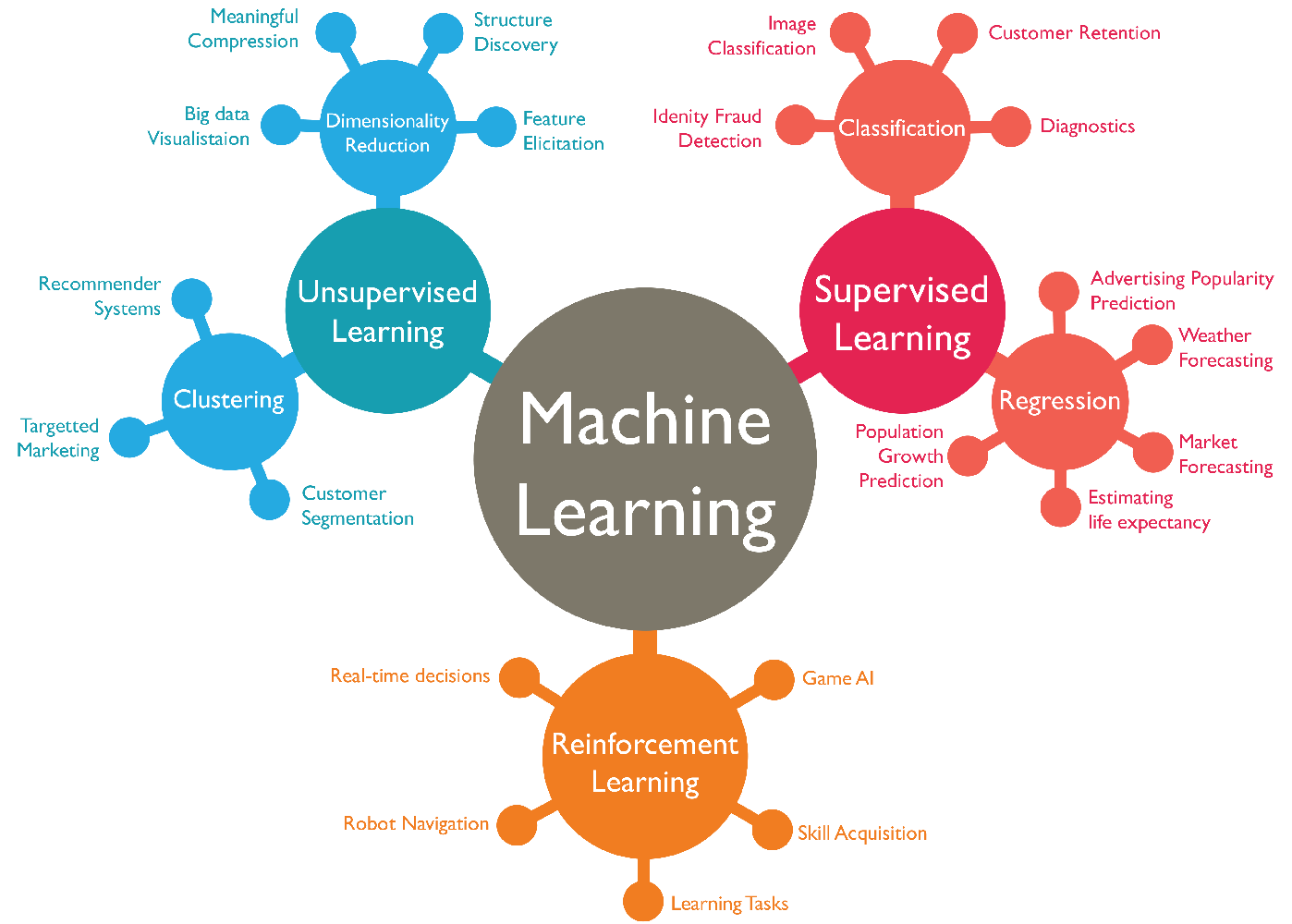
همانطور که از نام آن پیداست، در یادگیری نظارت‌یافته، سرپرستی وجود دارد تا به الگوریتم یادگیری این بینش را بدهد که یک عمل یا تصمیم تا چه حدی خوب یا بد است. در روش‌های یادگیری نظارت یافته، مجموعه داده‌ها کاملاً رده‌بندی شده‌اند و روش یادگیری می‌تواند بررسی کند که یک عمل خاص صحیح یا نادرست است و همچنین میزان صحت آن چقدر است. الگوریتم‌های یادگیری ماشین محبوب نظارت‌یافته عبارتنداز: ماشین بردار پشتیبان[[20]](#footnote-20)، جنگل تصادفی[[21]](#footnote-21) و شبکه عصبی[[22]](#footnote-22).

#### ***2-2-4-2 یادگیری ماشین بدون نظارت***

در این نوع از یادگیری، مجموعه داده‌ها دارای برچسب نیستند. این بدان معنی است که الگوریتم باید برچسب‌ها را پیدا کرده و آن‌ها را تعریف کند. چنین الگوریتم‌هایی نیاز به یادگیری ساختار مجموعه داده‌ها و رابطه بین ویژگی‌ها دارند. الگوریتم‌های یادگیری ماشین محبوب نظارت‌یافته عبارتنداز: خوشه‌بندی [[23]](#footnote-23)K-means و شبکه‌های عصبی خودسازمان‌یافته[[24]](#footnote-24).

#### **2-2-4-3 یادگیری ماشین تقویتی**

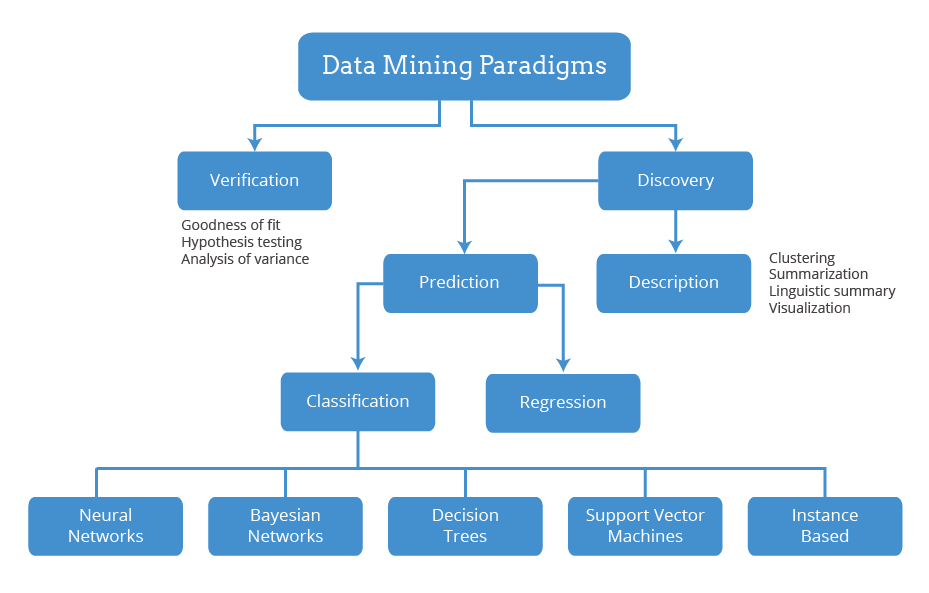
در یادگیری تقویتی، الگوریتم یادگیری در صورت انجام یک عمل صحیح پاداش دریافت می‌کند و در صورت انجام یک عمل اشتباه مجازات می‌شود. این نوع یادگیری مقدمه‌ای بر تکنیک‌های یادگیری ماشین تکاملی به شیوه یادگیری موجودات از طریق پاداش و مجازات را شبیه‌سازی می‌کند. برخی از نمونه‌های یادگیری تقویتی عبارتنداز: Q learning، [[25]](#footnote-25)DQN و DDPG[[26]](#footnote-26)( میرجلیلی و همکاران[1]).



شکل 2-2 کلاس‌های مختلف یادگیری ماشین

## 2-3 روش‌های داده‌کاوی

همانطور كه در شكل 2-2 مشاهده مي‌كنيد الگوي داده‌كاوي را مي‌توان به دو دسته‌ي كلي صحه‌گذاري و كشف تقسيم كرد. آن‌چه كه امروزه به عنوان روش‌هاي داده‌كاوي شناخته مي‌شود دسته‌ي دوم يعني كشف مي‌باشد. اين دسته خود به دو شاخه‌ي كلي پيش‌بيني كننده و توصيف‌كننده تقسيم مي‌شود. شاخه‌ي پيش‌بيني كننده خود به دو دسته‌ي طبقه‌بندي و رگرسيون تقسيم مي‌شود. از مهم‌ترين روش‌هايي كه در شاخه‌ي طبقه‌بندي قرار مي‌گيرند مي‌توان به شبكه‌هاي عصبي، شبكه‌هاي بيزين، درخت تصميم و ماشين‌هاي بردار پشتيبان را نام برد. شاخه‌ي توصيف‌كننده نيز خود به شاخه‌هايي همچون خوشه‌بندي، خلاصه‌سازي و مصورسازي تقسيم مي‌شود.



شكل 2-3 دسته‌بندي روش‌هاي مختلف داده‌كاوي

## 2-4 تحلیل داده‌های مالی ازدیدگاه‌های گوناگون

### **2-4-1 سير زماني تحقيق**

نظریه‌هاي متفاوتی در خصوص ارزیابی و پیش‌بینی بورس در بازارهاي سازمان‌یافته مطرح شده است. در اوایل قرن بیستم، گروهی از متخصصان صاحب تجربه در ارزیابی اوراق بهادار اعتقاد راسخ بر این امر داشتند که می‌توان از طریق مطالعه و تجزیه ‌و تحلیل روند تاریخی تغییرات قیمت سهام، تصویري را براي پیش‌بینی قیمت آینده سهام ارائه نمود. مطالعات علمی‌تر با تأکید بر شناسایی دقیق رفتار قیمت سهام، گرایش به سمت مدل‌هاي ارزشیابی قیمت سهام را به وجود آورد. در ابتدا نظریه‌ی گام‌هاي تصادفی 3 به‌ عنوان یک شروع در تعیین رفتار قیمت سهام مطرح شد. سپس به ویژگی‌ها و ساختار بازار سرمایه توجه شد که نتیجه‌ی این مطالعات و بررسی‌ها منجر به فرضیه‌ی بازار کارآي سرمایه شد.

این فرضیه به دلیل ترکیب خاص آن، مورد توجه محافل علمی قرار گرفت. در بازار کارآي سرمایه، اعتقاد بر این است که قیمت سهام انعکاسی از اطلاعات جاري مربوط به آن سهم است و تغییرات قیمت سهام داراي الگوي خاص قابل پیش‌بینی نیست. نظریات مطرح شده تا دهه‌ي 1980 میلادي به‌خوبی تعیین‌کننده‌ی رفتار قیمت سهام در بازار بودند تا اینکه تحولات بازار سهام نیویورك در سال 1987 میلادي، اعتبار فرضیات بازار کارآي سرمایه و مدل‌هایی نظیر تصادفی بودن قیمت‌ها را به‌شدت زیر سؤال برد. در دهه‌ی 1990 میلادي و بعد از آن، بیشتر توجه متخصصان به یک رفتار آشوبگرانه همراه بانظم معطوف شد و تلاش در جهت طراحی مدل‌هاي غیرخطی به‌منظور پیش‌بینی قیمت سهام اهمیت روزافزونی یافت.

با این نظریات، ازجمله تکنیک‌هایی که اهمیت بالایی یافتند، سیستم‌هاي هوشمند بودند؛ زیرا با فرض خطی بودن ساختار بازار، به‌آسانی می‌توان بسیاري از مدل‌ها را طراحی نمود. با این‌وجود، بسیار سخت است که بتوان رفتار مجموعه‌هاي پیچیده‌ای نظیر بازار سرمایه در یک مجموعه‌ي اقتصادي مدرن را به‌طور کامل در یک مجموعه معادلات ساده و خطی نشان داد. مزیت عمده‌ي سیستم‌هاي هوشمند نظیر شبکه‌هاي عصبی مصنوعی و شبکه‌هاي عصبی فازي، در مدل‌سازی و پیش‌بینی مجموعه‌هاي نامنظم و غیرخطی است. ابزار دیگري نظیر الگوریتم ژنتیک نیز ازنظر بسیاري از محققان می‌تواند در کاهش زمان به جواب رسیدن و حتی بهینه‌سازي پیش‌بینی‌ها در شبکه‌هاي عصبی مصنوعی و شبکه‌هاي عصبی فازي مثمر ثمر باشد.

### **2-4-2 مقالات انجام شده**

پیش‌بینی روند و قیمت دارایی‌های مالی یکی از موضوعات جذاب برای محققین در سال‌های اخیر بوده است. بازار مالی فارکس یکی از این بازار‌های مالی است که پژوهش‌های متعددی در بستر آن انجام شده است. از جمله‌ی این تحقیقات می‌توان به کار (باشر و فخر[2]) در سال 2011 اشاره کرد. آن‌ها در این تحقیق از پنجره زمانی چندگانه تحلیل فنی فارکس و پردازش سیگنال ویژگی‌ها جهت پیش‌بینی سریع نرخ روند روزانه بازار استفاده کردند و در کار آن‌ها پیش‌بینی به عنوان یک مسئله طبقه‌بندی باینری مطرح می شود. آن‌ها برای استخراج ویژگی‌های مورد نیاز در مدل طبقه‌بندی خود از پنج تکنیک و برای انتخاب ویژگی‌ها از دو تکنیک ماشین بردار پشتیبان و درخت بگینگ[[27]](#footnote-27) استفاده کردند. در پژوهشی دیگر (نصیرطوسی و همکاران[3]) به پیش‌بینی بازار فارکس با در نظرگیری اخبار و احساسات معامله‌گر پرداختند. آن‌ها در کار خود به بررسی چالشی که در این زمینه وجود داشت یعنی دسترسی به داده‌های بنیادی پنهان در متن‌های خبری غیرساختاریافته پرداختند. از جمله نوآوری‌هایی که در این پژوهش مشاهده می‌شود می‌توان به سه مورد که عبارتند از: انتخاب استراتژیک ویژگی ها با یک روش جدید و ابتکاری،ارائه یک الگوریتم کاهش ویژگی ها به نام کاهش ویژگی مبتنی بر هدف[[28]](#footnote-28) و ارائه یک روش وزن‌دهی احساسات جدید به نام مجموع امتیازات[[29]](#footnote-29) اشاره کرد. در نهایت آن‌ها در نتیجه‌ی کار خود نشان دادند که بین عناوین اخبار و حرکات قیمت جفت ارز رابطه‌ی امیدوارکننده‌ای وجود دارد. در یکی از پژوهش‌های دیگر که دارای ارجاع به مقاله بالایی نیز می‌باشد(یائو و لیم‌تان[4]) به پیش‌بینی فنی بازار فارکس با استفاده از شبکه‌های عصبی پرداختند. آن‌ها در نتیجه‌گیری‌های کار خود به این نکته اشاره کرده‌اند که دقت پیش‌بینی در تمامی جفت‌ارز‌ها به جز دلار-ین بسیار قابل قبول می‌باشد. آن‌ها در نتیجه‌گیری خود اشاره کردند که بازار ین بزرگتر و با ارزش تر از بازارهای دلار استرالیا، فرانک سوئیس و پوند انگلیس است. همچنین معامله‌گران بازار ین ممکن است بیشتر به تجزیه و تحلیل فنی بستگی داشته باشند و بعد از ظهور هر نشانه ای به سرعت عمل کنند. از این رو، تجزیه و تحلیل فنی ممکن است ابزار مناسبی برای پیش‌بینی روندهای ین نباشد ، زیرا همه‌ی معامله‌گران از معنای سیگنال‌های فنی آگاه هستند.

در اکثر مقالات مورد بررسی جهت ارائه پیش‌بینی از ابزارهای تحلیل فنی استفاده شده است و داده‌های تحلیل بنیادی در نظر گرفته نشده اند. حال آن‌که استفاده از این داده‌ها می‌توانند به نتایج قابل قبولی منجر شوند. (نصیرطوسی و همکاران[5]) در پژوهش خود این موضوع را مورد بررسی قرار دادند و در کار خود، تلاش کرده‌اند تا امکان استفاده از داده های بنیادی برای پیش‌بینی حرکت قیمت ارز در بازار فارکس را بررسی کنند. این نوع از پیش بینی در رابطه با بررسی تقاضا بسیار متداول است؛ با این حال، تجزیه و تحلیل فنی رویکردی است که به طور گسترده در تحقیقات در این زمینه مورد بررسی قرار گرفته‌ است. روش پیشنهادی در این پژوهش که به بهره‌برداری از شبکه‌های عصبی منتهی می‌شود، با آزمایش‌های انجام شده اثبات می‌گردد. نتایج آزمایش‌های انجام شده در این مقاله نیز نشانگر قابل قبول بودن مدل ارائه شده در تعیین حرکت ارزی از طریق روش پیشنهادی و با استفاده از ورودی مشخص شده بود. همچنین آن‌ها در نتیجه‌گیری کار خود اظهار داشتند که روش ارائه شده توسط آن‌ها علاوه بر شناسایی برخی از داده‌های یبنیادی که می‌تواند برای چنین پیش بینی‌هایی استفاده شود و یک متدولوژی را پیشنهاد ‌کند، همچنین می‌تواند از طریق آزمایش های انجام شده نشان دهد که اگرچه مجموعه ای از داده‌های بنیادی ممکن است نشان دهنده حرکت قیمت به خودی خود نباشد ، اما ممکن است در ترکیب با سایر مجموعه‌های داده هایی از این دست، به تعیین چنین نشانه‌هایی کمک کند.

# فصل سوم: پیش‌پردازش داده‌ها

## 3-1 مقدمه

شروع هر نوع کار و عملیاتی در مرحله اول، دارای یک سری مقدمات و پیش‌نیازها است. داده‌کاوی نیز از این قانون مستثنی نبوده و نیازمند آماده‌سازی و پردازش‌های مقدماتی است. در علم داده‌کاوی، تمامی داده‌هایی که برای هدف مورد نظر استفاده خواهند شد، باید پیش از شروع پردازش با استفاده از روش‌هایی، آماده و تنظیم و یا به اصطلاح پیش‌پردازش[[30]](#footnote-30) شوند. پیش‌پردازش نقشی اساسی در روند پردازش داده‌ها و نتایج حاصل از آن‌ها ایفا می‌کند. برای پیش پردازش داده ها مراحل و ابزارهای مختلفی وجود دارند. برخی از مهم‌ترین مواردی که طی فرایند پیش‌پردازش داده‌ها باید به آن‌ها پرداخته شود عبارتند از:

داده‌های ناموجود[[31]](#footnote-31)

داده‌های پرت[[32]](#footnote-32)

نرمال‌سازی داده‌ها[[33]](#footnote-33)

## 3-2 پیش پردازش داده‌ها

در این پژوهش نیز در ابتدا به آماده‌سازی و پردازش‌های مقدماتی داده‌ها می‌پردازیم. گام‌های پیش‌پردازش اعمال شده به شرح زیر می‌باشد:

1. بدلیل اینکه هدف این پژوهش پیش‌بینی ستون متغیر Target1 برای روز آینده به عنوان متغیر پیش‌بینی شونده است در نتیجه ابتدا دیتای مربوط به این متغیر را یک روز به سمت جلو منتقل می‌کنیم. به بیان دیگر دیتایی که بعنوان متغیر Target1 مورد نیاز است از روز دوم تا روز آخر در نظر گرفته شده است.
2. در قسمت بعد ستون تاریخ و ستون متغیر Target2 که شامل رنگ کندل(سبز و قرمز) می‌باشد حذف می‌گردد و به جای آن اعداد 1(نماینده رنگ سبز و حرکت صعودی) و 0(نماینده رنگ قرمز و حرکت نزولی) قرار داده می‌شوند.
3. در گام بعد(خط 12) از data.describe جهت نرمال استاندارد کردن داده‌ها استفاده شده است.
4. در مرحله بعد به دلیل اینکه داده‌های مربوط به قیمت داده‌های ايستا نبوده و میانگین و واریانسشان در طول زمان تغییر می‌کند می‌بایست به یک سری زمانی دیگر تبدیل شوند که میانگین و واریانسشان ثابت شود. به همین دلیل برای اینکه بتوانیم از قیمت به درستی استفاده کنیم از بازدهی به صورت جایگزین استفاده می‌کنیم. برای این کار 2 نوع بازدهی تعریف شده است که عبارتند از: بازدهی ساده که میزان درصد تغییر قیمت در روز را نشان می‌دهد و بازدهی لگاریتمی که تقسیم مقادیر لگاریتمی داده‌ها را جایگزین می‌کند که در این پژوهش(در حلقه‌ی for خط 26 ام) از بازدهی لوگاریتمی استفاده شده است.
5. برخی از ویژگی‌های دیگر(شامل macd، signal، rsi، mavg و ...) که از جنس قیمت نمی‌باشند هم به صورت نرمال استاندارد و هم به صورت لگاریتمی به داده‌ها اضافه شده‌اند.
6. برخی از داده‌ها که در هنگام محاسبه لگاریتم مقادیر مثبت یا منفی بی‌نهایت را اختیار می‌کردند در خط 72 حذف شدند.

# فصل چهارم: شناسایی ویژگی‌های مورد نظر جهت داده‌کاوی

## 4-1 مقدمه

در مواقعی که بحث کار عملی بر روی داده‌ها پیش می‌آید و از مباحث تئوری فاصله می‌گیریم، شاید مهم‌ترین بخش برای عملیات داده‌کاویْ عملیات انتخاب ویژگی[[34]](#footnote-34) است. در مباحث آکادمیک معمولا ویژگی‌ها در مسئله در اختیار کاربران قرار دارند ولی در مباحث عملی یک متخصص علوم‌داده بایستی خود ویژگی‌های مورد نیاز را از میان داده‌ها استخراج کند. حتی ممکن است نیاز باشد به دنبال ساخت دیتاست جدید بگردد و داده‌ها را جمع آوری کند. هدف از انتخاب ویژگی، بهبود عملکرد پیش‌بینی، ارائه پیش‌بینی سریع‌تر و مقرون‌به‌صرفه‌تر و ارائه‌ی درک بهتر از روند اطلاعات تولید شده است.

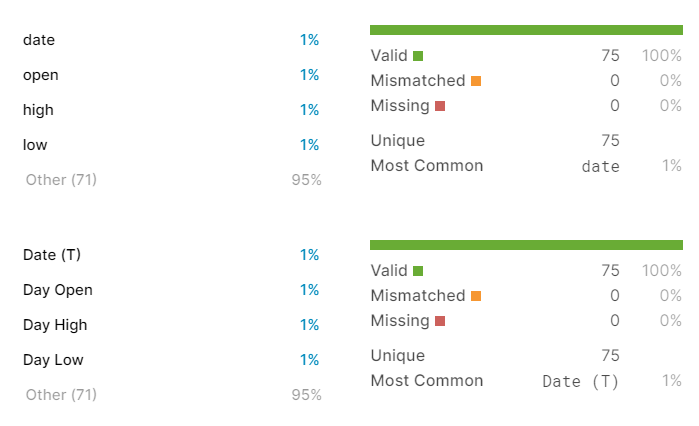
در این فصل ابتدا به معرفی ویژگی‌های موجود در دیتاست پژوهش می‌پردازیم. سپس در بخش بعد ویژگی مورد نظر جهت ایجاد مدل و بهبود عملکرد پیش‌بینی را شناسایی می‌کنیم.

## 4-2 معرفی ویژگی‌های موجود

ویژگی‌های موجود در دیتاست این پژوهش به شرح جدول زیر می‌باشند. همچنين در شكل 4-1 مي‌توانيد اطلاعات مربوط به اين ويژگي‌ها كه در وبسايت [www.kaggle.com](http://www.kaggle.com) موجود است را مشاهده فرماييد.

جدول 4-1 ويژگي‌هاي موجود در مجموعه داده‌ي مورد بررسي پژوهش

|  |  |  |
| --- | --- | --- |
| Number | Name | Description |
| **1** | date | Date (T) |
| **2** | open | Day Open |
| **3** | high | Day High |
| **4** | low | Day Low |
| **5** | close | Day Close |
| **6** | open1 | T-1 Open |
| **7** | high1 | T-1 High |
| **8** | low1 | T-1 Low |
| **9** | close1 | T-1 Close |
| **10** | open2 | T-2 Open |
| **11** | high2 | T-2 High |
| **12** | low2 | T-2 Low |
| **13** | close2 | T-2 Close |
| **14** | open3 | T-3 Open |
| **15** | high3 | T-3 High |
| **16** | low3 | T-3 Low |
| **17** | close3 | T-3 Close |
| **18** | open4 | T-4 Open |
| **19** | high4 | T-4 High |
| **20** | low4 | T-4 Low |
| **21** | close4 | T-4 Close |
| **22** | open5 | T-5 Open |
| **23** | high5 | T-5 High |
| **24** | low5 | T-5 Low |
| **25** | close5 | T-5 Close |
| **26** | macd0 | MACD |
| **27** | signal0 | MACD Signal |
| **28** | diff0 | MACD Diff |
| **29** | macd1 | T-1 MACD |
| **30** | signal1 | T-1 MACD Signal |
| **31** | diff1 | T-1 MACD Diff |
| **32** | macd2 | T-2 MACD |
| **33** | signal2 | T-2 MACD Signal |
| **34** | diff2 | T-2 MACD Diff |
| **35** | macd3 | T-3 MACD |
| **36** | signal3 | T-3 MACD Signal |
| **37** | diff3 | T-3 MACD Diff |
| **38** | macd4 | T-4 MACD |
| **39** | signal4 | T-4 MACD Signal |
| **40** | diff4 | T-4 MACD Diff |
| **41** | macd5 | T-5 MACD |
| **42** | signal5 | T-5 MACD Signal |
| **43** | diff5 | T-5 MACD Diff |
| **44** | rsi0 | RSI |
| **45** | rsi1 | T-1 RSI |
| **46** | rsi2 | T-2 RSI |
| **47** | rsi3 | T-3 RSI |
| **48** | rsi4 | T-4 RSI |
| **49** | rsi5 | T-5 RSI |
| **50** | dn0 | Lower Bollinger Band |
| **51** | mavg0 | Middle Moving Average |
| **52** | up0 | Upper Bollinger Band |
| **53** | pctB0 | Percentage B |
| **54** | dn1 | T-1 Lower Bollinger Band |
| **55** | mavg1 | T-1 Middle Moving Average |
| **56** | up1 | T-1 Upper Bollinger Band |
| **57** | pctB1 | T-1 Percentage B |
| **58** | dn2 | T-2 Lower Bollinger Band |
| **59** | mavg2 | T-2 Middle Moving Average |
| **60** | up2 | T-2 Upper Bollinger Band |
| **61** | pctB2 | T-2 Percentage B |
| **62** | dn3 | T-3 Lower Bollinger Band |
| **63** | mavg3 | T-3 Middle Moving Average |
| **64** | up3 | T-3 Upper Bollinger Band |
| **65** | pctB3 | T-3 Percentage B |
| **66** | dn4 | T-4 Lower Bollinger Band |
| **67** | mavg4 | T-4 Middle Moving Average |
| **68** | up4 | T-4 Upper Bollinger Band |
| **69** | pctB4 | T-4 Percentage B |
| **70** | dn5 | T-5 Lower Bollinger Band |
| **71** | mavg5 | T-5 Middle Moving Average |
| **72** | up5 | T-5 Upper Bollinger Band |
| **73** | pctB5 | T-5 Percentage B |
| **74** | target1 | Open and Close Difference in pips |
| **75** | target2 | Color of Candle Stick - Green, Red, Black |



شكل 4-1 اطلاعات ويژگي‌هاي مورد استفاده در پژوهش

# فصل پنجم: انتخاب ویژگی‌های مطلوب

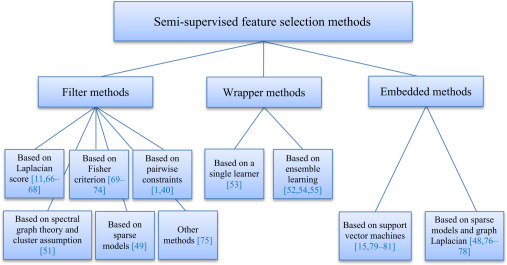
## 5-1 متدولوژي‌هاي موجود جهت انتخاب ويژگي

روش‌های انتخاب ویژگی[[35]](#footnote-35) به منظور مواجهه با داده‌های ابعاد بالا، به مولفه‌ای جدایی ناپذیر از فرآیند یادگیری مبدل شده‌اند. به طور كلي همانطور كه در شكل 5-1 مشخص است روش‌هاي موجود جهت انتخاب ويژگي‌ها را مي‌توان به سه دسته تقسيم كرد كه عبارتند از:

«فیلترها[[36]](#footnote-36)» : اين نوع از روش‌ها بر ویژگی‌های کلی مجموعه داده آموزش تکیه دارند و فرآیند انتخاب ویژگی را به عنوان یک گام پیش پردازش با استقلال از الگوریتم استقرایی انجام می‌دهند. مزیت این مدل‌ها هزینه محاسباتی پایین و توانایی تعمیم خوب آن‌ها محسوب می‌شود.

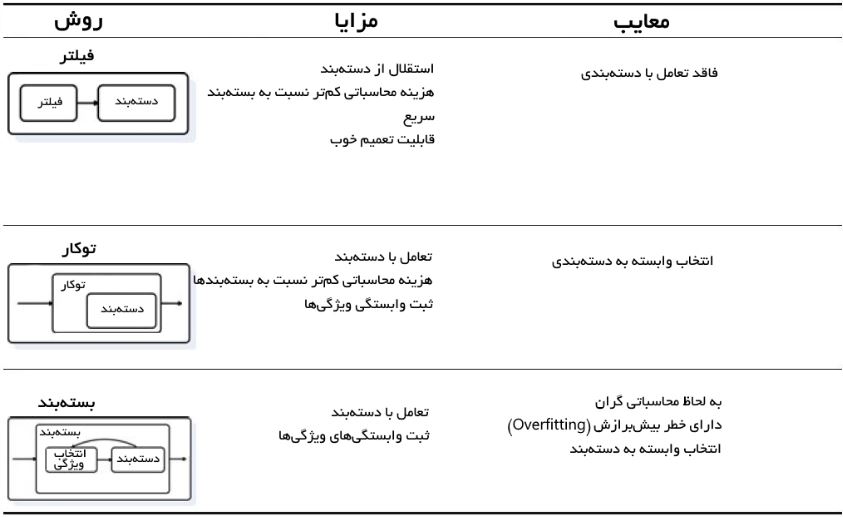
«بسته‌بندها[[37]](#footnote-37)» : شامل یک الگوریتم یادگیری به عنوان جعبه سیاه هستند و از کارایی پیش‌بینی آن برای ارزیابی مفید بودن زیرمجموعه‌ای از متغیرها استفاده می‌کنند. به عبارت دیگر، الگوریتم انتخاب ویژگی از روش یادگیری به عنوان یک زیرمجموعه با بار محاسباتی استفاده می‌کند که از فراخوانی الگوریتم برای ارزیابی هر زیرمجموعه از ویژگی‌ها نشات می‌گیرد. با این حال، این تعامل با دسته‌بند منجر به نتایج کارایی بهتری نسبت به فیلترها می‌شود.

«روش‌های توکار[[38]](#footnote-38)» : انتخاب ویژگی را در فرآیند آموزش انجام می‌دهند و معمولا برای ماشین‌های یادگیری خاصی مورد استفاده قرار می‌گیرند. در این روش‌ها، جست‌و‌جو برای یک زیرمجموعه بهینه از ویژگی‌ها در مرحله ساخت دسته‌بند انجام می‌شود و می‌توان آن را به عنوان جست‌و‌جویی در فضای ترکیبی از زیر مجموعه‌ها و فرضیه‌ها دید. این روش‌ها قادر به ثبت وابستگی‌ها با هزینه‌های محاسباتی پایین‌تر نسبت به بسته‌بندها هستند.



شكل 5-1 روش‌هاي انتخاب ويژگي

در شكل زیر خلاصه‌ای از سه روش انتخاب ویژگی معرفی شده در بالا آمده و برجسته‌ترین مزایا و معایب آن‌ها را بیان شده است.



شكل 5-2 مزايا و معايب روش‌هاي انتخاب ويژگي

## 5-2 انتخاب ويژگي‌هاي با بيشترين اثرگذاري

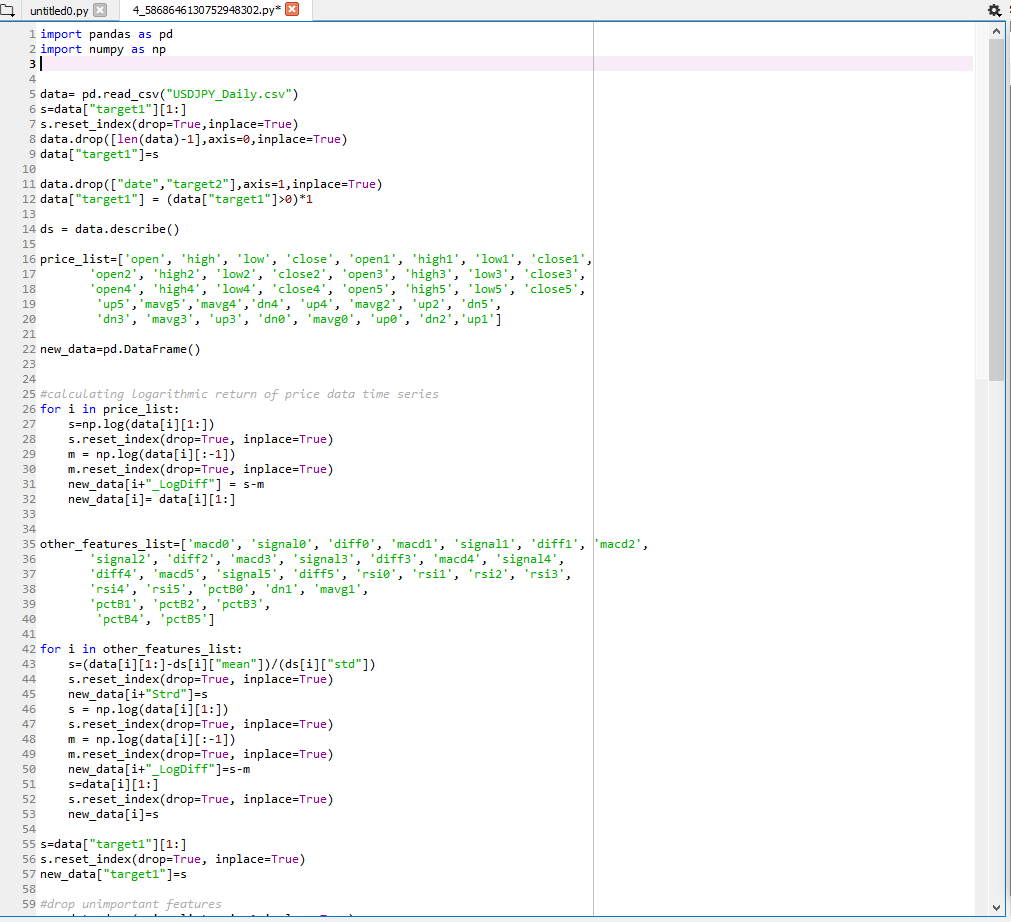
در اين قسمت با استفاده از توانايي الگوريتم طبقه‌بندي درخت تصميم در شناسايي درجه اهميت پارامترها(توانايي در توضيح نتايج توسط پارامتر مورد نظر) آن‌ها را مرتب مي‌كنيم كه در نهايت پارامترهايي با درجه اهميت 0 را از مدل خارج مي‌كنيم.

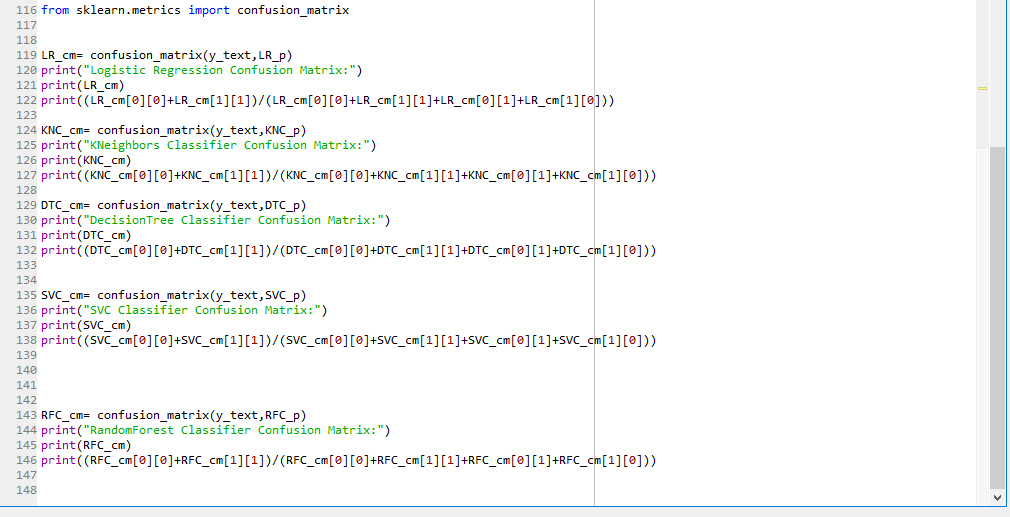
جدول 5-1 مرتب‌سازي ويژگي‌هاي موجود در مسئله بر اساس ميزان اهميت

|  |  |
| --- | --- |
| Parameters | FI |
| high\_LogDiff | 0.280191 |
| pctB0\_LogDiff | 0.070256 |
| pctB2Strd | 0.043798 |
| open1\_LogDiff | 0.039968 |
| rsi0\_LogDiff | 0.03328 |
| rsi4Strd | 0.031478 |
| dn1Strd | 0.029254 |
| diff2Strd | 0.024525 |
| high4\_LogDiff | 0.024078 |
| low\_LogDiff | 0.022293 |
| mavg2\_LogDiff | 0.021997 |
| diff3\_LogDiff | 0.020247 |
| diff4Strd | 0.020225 |
| pctB1Strd | 0.0196 |
| rsi2Strd | 0.019443 |
| low1\_LogDiff | 0.019263 |
| close5\_LogDiff | 0.017952 |
| high2\_LogDiff | 0.015479 |
| diff0\_LogDiff | 0.015399 |
| pctB5\_LogDiff | 0.014492 |
| mavg4\_LogDiff | 0.014447 |
| up0\_LogDiff | 0.014187 |
| pctB4\_LogDiff | 0.013544 |
| pctB4Strd | 0.01306 |
| dn1\_LogDiff | 0.012641 |
| close\_LogDiff | 0.012304 |
| diff5Strd | 0.012183 |
| open3\_LogDiff | 0.011507 |
| dn0\_LogDiff | 0.011454 |
| diff5\_LogDiff | 0.008839 |
| ... | ... |
| low2\_LogDiff | 0 |
| close1\_LogDiff | 0 |
| high1\_LogDiff | 0 |
| pctB5Strd | 0 |
| open5\_LogDiff | 0 |
| high5\_LogDiff | 0 |
| signal0\_LogDiff | 0 |
| low5\_LogDiff | 0 |
| diff0Strd | 0 |
| mavg0\_LogDiff | 0 |
| macd1\_LogDiff | 0 |
| diff1Strd | 0 |
| signal2Strd | 0 |
| dn3\_LogDiff | 0 |
| diff2\_LogDiff | 0 |
| macd3Strd | 0 |
| signal3\_LogDiff | 0 |
| dn5\_LogDiff | 0 |
| signal4Strd | 0 |
| signal4\_LogDiff | 0 |
| up2\_LogDiff | 0 |
| diff4\_LogDiff | 0 |
| macd5Strd | 0 |
| macd5\_LogDiff | 0 |
| signal5Strd | 0 |
| signal5\_LogDiff | 0 |
| dn4\_LogDiff | 0 |
| mavg5\_LogDiff | 0 |
| rsi0Strd | 0 |
| open\_LogDiff | 0 |

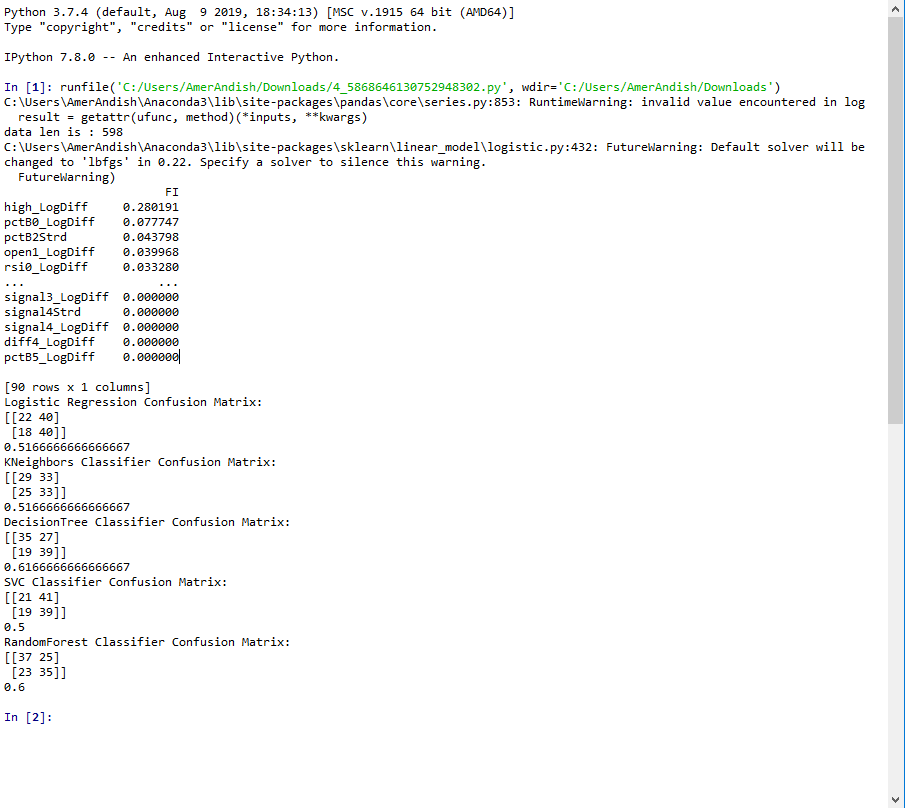
# فصل ششم: داده‌کاوی و شناسایی الگوهای پنهان در داده‌ها، و ارائه نهایی مدل‌هاي طبقه‌بندي

در اين بخش كليه‌ي مراحل تشريح شده در فصل‌هاي پيشين و همچنين برخي از مدل‌هاي پركاربرد در زمينه طبقه‌بندي بر روي ديتاست معرفي شده در نرم‌افزار پايتون پياده‌سازي شده است.





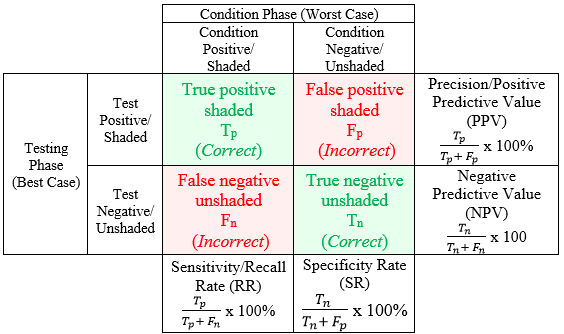
شكل 7-1 پياده‌سازي مراحل تشريح شده در نرم افزار پايتون



شكل 7-2 نتايج حاصل از مدل

نتايج حاصل از مدل‌هاي اعمال شده بوسيله ماتريس درهم‌ريختگي[[39]](#footnote-39) نمايش داده مي‌شود. اين ماتريس كه در شكل 7-1 نشان داده شده است از 4 قسمت تشكيل شده است كه عبارتند از:

* بخش شمال غربي: داده‌هاي مثبتي كه به درستي مثبت پيش‌بيني شده‌اند. كه در اين تحقيق شامل روزهايي است كه روند حركت جفت‌ارز صعودي بوده و ومدل نيز به درستي صعودي پيش‌بيني كرده است.
* بخش شمال شرقي: داده‌هاي منفي كه به اشتباه مثبت پيش‌بيني شده‌اند. كه در اين تحقيق شامل روزهايي است كه روند حركت جفت‌ارز نزولي بوده و ومدل به اشتباه صعودي پيش‌بيني كرده است.
* بخش جنوب غربي: داده‌هاي مثبتي كه به اشتباه منفي پيش‌بيني شده‌اند. كه در اين تحقيق شامل روزهايي است كه روند حركت جفت‌ارز صعودي بوده و ومدل به اشتباه نزولي پيش‌بيني كرده است.
* بخش جنوب شرقي: داده‌هاي منفي كه به درستي منفي پيش‌بيني شده‌اند. كه در اين تحقيق شامل روزهايي است كه روند حركت جفت‌ارز نزولي بوده و ومدل نيز به درستي نزولي پيش‌بيني كرده است.



شكل 7-3 ماتريس درهم‌ريختگي

همچنين با استفاده از مقادير موجود در اين ماتريس مي‌توان مقادير شاخص‌هاي صحت[[40]](#footnote-40)، حساسيت[[41]](#footnote-41)، وضوح[[42]](#footnote-42) و در نهايت ميزان دقت[[43]](#footnote-43) مدل را محاسبه كرد.

بنابر نتايج حاصل شده پس از پياده‌سازي مدل ميزان دقت مدل‌ها بر اساس ماتريس درهم‌ريختگي برابر است با:

# فصل هفتم: نتیجه‌گیری

در اين پژوهش به بررسي بازار جهاني فاركس و پيش‌بيني روند حركت يكي از جفت‌ارزهاي اين بازار يعني دلار-ين پرداخته شده است. در گام اول به بررسي كلي داده‌هاي موجود بوسيله نرم‌افزار R پرداختيم تا يك ديد كلي از وضعيت داده‌ها حاصل شود. در ادامه جهت دستيابي به نتيجه‌ي بهتر از مدل به آماده‌سازي داده‌ها پرداخته شد. پس از آن از بين پارامترهاي موجود برخي از پارامترها با درجه اهميت بالاتر براي استفاده در مدل‌هاي طبقه‌بندي گزينش شدند و در گام آخر نيز برخي از مدل‌هاي پركاربرد در طبقه‌بندي شامل رگرسيون لجستيك[[44]](#footnote-44)، درخت تصادفي، ماشين بردار پشتيبان، جنگل تصادفي و k همسايه[[45]](#footnote-45) براي طبقه‌بندي داده‌ها و پيش‌بيني روند حركت جفت-ارز مذكور در روز آينده مود بررسي قرار گرفتند كه نتيجه‌ي حاصل از روش‌هاي درخت تصميم و جنگل تصادفي از بالاترين دقت برخوردار بودند. البته لازم به ذكر است كه بدليل ماهيت داده‌هاي موجود و عدم ايستايي در داده‌ها پيش‌بيني آن‌ها بسيار دشوار است و در اكثر مقالات مورد بررسي نيز دقت مدل‌هاي حاصل از حد مشخصي(حدود 70%) تجاوز نكرده است كه اين موضوع خود حاكي از دشواري پيش‌بيني در چنين فضايي است و همين امر دليل خوبي براي نياز به تحقيقات بيشتر و گسترده تر در اين زمينه مي‌باشد.

# پیوست: کد نرم‌افزار و توضیح درباره نرم‌افزار مورد استفاده

## نرم‌افزار مورد استفاده جهت پياده سازي مدل

پایتون یک زبان برنامه نویسی قدرتمند سطح بالا، شئ گرا و حرفه‌ای می‌باشد که در حال گسترش روز‌افزون در جهان است. پایتون از جمله زبان‌های برنامه‌نویسی قدرتمندی است که در زمینه علم داده‌ها، یادگیری ماشینی، خودکارسازی سامانه‌ها، توسعه وب، واسط‌های برنامه‌نویسی و... به کار گرفته می‌شود. این زبان با قابلیت‌های فراوان و شگفت انگیزی که دارد تحولی در دنیای برنامه نویسی از توسعه برنامه‌های تحت وب تا ایجاد بازی‌های رایانه‌ای، بوجود آورده است. پایتون ابتدا در سال 1991 وارد دنیای برنامه نویسی شد و در سال‌های اخیر توجه برنامه نویسان را به خود جلب کرده و روز به روز بر طرفداران آن افزوده می‌شود. پایتون هنوز در ایران جایگاه خود را پیدا نکرده است اما آینده روشنی برای آن می‌توان تصورکرد زیرا این زبان کاربردهای فراوانی دارد و در بسیاری از سایت‌های بین‌المللی نیز مورد استفاده قرار گرفته است.

تعداد کلمات کلیدی پایتون کم، ساده و کاملا قابل درک است و این موضوع فهم و یادگیری آن را برای کاربران تازه‌کار بسیار ساده کرده است. در واقع این زبان پیچیدگی‌های معمول سایر زبان‌ها را ندارد و پس از برنامه نویسی، منطق آن کاملا قابل درک است. این زبان اپن سورس را می‌توان در زمان کوتاهی به خوبی یاد گرفت و بواسطه کتابخانه‌های گسترده‌ای که دارد از آن استفاده‌های فراوان کرد.

پایتون یک زبان اسکریپتی است به این منظور که کدهای آن در اکثر پلت فرم‌ها از جمله لینوکس، ویندوز، مکینتاش، سیستم‌های موبایل و حتی پلی‌استیشن قابل اجراست. این زبان به سبب قابلیت‌های فراوانی که دارد زبان مورد علاقه برنامه نویسان وب می‌باشد. شرکت‌های عظیمی مانند گوگل، یاهو، ناسا و … در سطح وسیعی در حال استفاده از پایتون هستند. از جمله كاربردهاي اين زبان محبوب مي‌توان به موارد زير اشاره كرد:

* خودکار‌سازی برنامه‌ها
* گسترش برنامه‌های تحت وب
* اسکریپت نویسی
* آنالیز اطلاعات
* توسعه اپلیکیشن‌های تحت وب

## كد نرم‌افزار

کد نرم افزار در فایل Dataminingproject.py به پژوهش ضمیمه شده است.

# منابع و مراجع

1. Mirjalili, S., H. Faris, and I. Aljarah, *Introduction to Evolutionary Machine Learning Techniques*, in *Evolutionary Machine Learning Techniques*. 2020, Springer. p. 1-7.

2. Baasher, A.A. and M.W. Fakhr. *Forex trend classification using machine learning techniques*. in *Proceedings of the 11th WSEAS international conference on Applied computer science*. 2011. World Scientific and Engineering Academy and Society (WSEAS).

3. Nassirtoussi, A.K., et al., *Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment.* 2015. **42**(1): p. 306-324.

4. Yao, J. and C.L.J.N. Tan, *A case study on using neural networks to perform technical forecasting of forex.* 2000. **34**(1-4): p. 79-98.

5. Nassirtoussi, A.K., T.Y. Wah, and D.N.C.J.A.J.o.B.M. Ling, *A novel FOREX prediction methodology based on fundamental data.* 2011. **5**(20): p. 8322.

1. Information Technology [↑](#footnote-ref-1)
2. Databases [↑](#footnote-ref-2)
3. E-commerce [↑](#footnote-ref-3)
4. Data mining [↑](#footnote-ref-4)
5. USDJPY [↑](#footnote-ref-5)
6. FOREX [↑](#footnote-ref-6)
7. USDJPY [↑](#footnote-ref-7)
8. FOREX [↑](#footnote-ref-8)
9. Data Mining [↑](#footnote-ref-9)
10. Machine Learning [↑](#footnote-ref-10)
11. Database Management [↑](#footnote-ref-11)
12. Data Visualization [↑](#footnote-ref-12)
13. signal processing [↑](#footnote-ref-13)
14. Data Fishing [↑](#footnote-ref-14)
15. Data Dredging [↑](#footnote-ref-15)
16. Knowledge Discovery From Data [↑](#footnote-ref-16)
17. Supervise Learning [↑](#footnote-ref-17)
18. Unsupervised Learning [↑](#footnote-ref-18)
19. Reinforcement Learning [↑](#footnote-ref-19)
20. Support vector machines [↑](#footnote-ref-20)
21. Random forest [↑](#footnote-ref-21)
22. neural networks [↑](#footnote-ref-22)
23. K-means clustering [↑](#footnote-ref-23)
24. Self-organizing Neural Networks [↑](#footnote-ref-24)
25. Deep Q Network [↑](#footnote-ref-25)
26. Deep Deterministic Policy Gradient [↑](#footnote-ref-26)
27. Bagging Trees [↑](#footnote-ref-27)
28. targetbased feature-reduction [↑](#footnote-ref-28)
29. SumScores [↑](#footnote-ref-29)
30. Preprocess [↑](#footnote-ref-30)
31. Missing Data [↑](#footnote-ref-31)
32. Outliers [↑](#footnote-ref-32)
33. Normalization [↑](#footnote-ref-33)
34. Feature selection [↑](#footnote-ref-34)
35. Feature Selection Methods [↑](#footnote-ref-35)
36. Filters [↑](#footnote-ref-36)
37. Wrappers [↑](#footnote-ref-37)
38. Embedded [↑](#footnote-ref-38)
39. Confusion matrix [↑](#footnote-ref-39)
40. precision [↑](#footnote-ref-40)
41. sensitivity [↑](#footnote-ref-41)
42. specificity [↑](#footnote-ref-42)
43. Accuracy [↑](#footnote-ref-43)
44. Logistic regression [↑](#footnote-ref-44)
45. K neighbors [↑](#footnote-ref-45)