



دانشگاه صنعتی شریف

دانشگاه صنعتی شریف

دانشکده مهندسی صنایع

پروژه درس داده کاوی

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

نگارش:

آرش سرایی ۹۶۲۰۶۸۵۲

نیلوفر مبصری ۹۷۲۰۸۸۵۵

استاد درس:

جناب آقای دکتر مجید خدمتی

نیمسال اول ۹۸-۹۹

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

فصل اول: مقدمه.....	۵
فصل دوم: مروری بر ادبیات.....	۷
۱-۲ مقدمه.....	۸
۲-۲ اصول و تعاریف داده کاوی.....	۸
۱-۲-۲ تاریخچه داده کاوی.....	۸
۲-۲-۲ فرآیند داده کاوی.....	۹
۳-۲-۲ کاربردهای داده کاوی.....	۹
۴-۲-۲ یادگیری ماشین.....	۱۰
۳-۲ روش های داده کاوی.....	۱۱
۴-۲ تحلیل داده های مالی از دیدگاه های گوناگون.....	۱۲
۱-۴-۲ سیر زمانی تحقیق.....	۱۲
۲-۴-۲ مقالات انجام شده.....	۱۳
فصل سوم: پیش پردازش داده ها.....	۱۵
۱-۳ مقدمه.....	۱۶
۲-۳ پیش پردازش داده ها.....	۱۶
فصل چهارم: شناسایی ویژگی های مورد نظر جهت داده کاوی.....	۱۷
۱-۴ مقدمه.....	۱۸
۲-۴ معرفی ویژگی های موجود.....	۱۸
فصل پنجم: انتخاب ویژگی های مطلوب.....	۲۱
۱-۵ متدولوژی های موجود جهت انتخاب ویژگی.....	۲۲
۲-۵ انتخاب ویژگی های با بیشترین اثرگذاری.....	۲۳
فصل ششم: داده کاوی و شناسایی الگوهای پنهان در داده ها، و ارائه نهایی مدل های طبقه بندی.....	۲۶
فصل هفتم: نتیجه گیری.....	۳۲
پیوست: کد نرم افزار و توضیح درباره نرم افزار مورد استفاده.....	۳۴
❖ نرم افزار مورد استفاده جهت پیاده سازی مدل.....	۳۴
❖ کد نرم افزار.....	۳۴

فهرست اشکال و جداول

- شکل ۱-۲ گام‌های فرآیند داده‌کاوی ۹
- شکل ۲-۲ کلاس‌های مختلف یادگیری ماشین ۱۱
- شکل ۳-۲ دسته‌بندی روش‌های مختلف داده‌کاوی ۱۲
- جدول ۱-۴ ویژگی‌های موجود در مجموعه داده‌ی مورد بررسی پژوهش ۱۸
- شکل ۱-۴ اطلاعات ویژگی‌های مورد استفاده در پژوهش ۲۰
- شکل ۱-۵ روش‌های انتخاب ویژگی ۲۲
- شکل ۲-۵ مزایا و معایب روش‌های انتخاب ویژگی ۲۳
- جدول ۱-۵ مرتب‌سازی ویژگی‌های موجود در مسئله بر اساس میزان اهمیت ۲۳
- شکل ۱-۷ پیاده‌سازی مراحل تشریح شده در نرم افزار پایتون ۲۸
- شکل ۲-۷ نتایج حاصل از مدل ۲۹
- شکل ۳-۷ ماتریس درهم‌ریختگی ۳۰

فصل اول: مقدمه

با پیشرفت سریع فناوری اطلاعات^۱، بشر شاهد یک رشد انفجاری در تولید داده و ظرفیت‌های گردآوری و ذخیره‌سازی آن در دامنه‌های گوناگون بوده است. در جهان کسب‌وکار، پایگاه‌داده‌های^۲ بسیار بزرگی برای تراکنش‌های تجاری وجود دارند که توسط خرده‌فروشان و یا در تجارت الکترونیک^۳ 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^۴ 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^۵ در بازار فارکس^۶ 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^۷ در بازار فارکس^۸ 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.

¹ Information Technology

² Databases

³ E-commerce

⁴ Data mining

⁵ USDJPY

⁶ FOREX

⁷ USDJPY

⁸ FOREX

فصل دوم: مروری بر ادبیات

۲-۱ 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 می‌کنیم که روزانه مقادیر عظیمی از داده‌ها در آن جمع‌آوری می‌شوند. تجزیه و تحلیل چنین داده‌هایی یک نیاز مهم و ضروری است (دیتا ماینینگ-کانسپتز اند تکنیکز).

۲-۲ اصول و تعاریف داده‌کاوی

ما در جهانی زندگی می‌کنیم که روزانه مقادیر عظیمی از داده‌ها در آن جمع‌آوری می‌شوند. تجزیه و تحلیل چنین داده‌هایی یک نیاز مهم و ضروری است (دیتا ماینینگ-کانسپتز اند تکنیکز). داده‌کاوی^۹ 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^{۱۰}، مدیریت پایگاه داده^{۱۱} و بصری‌سازی داده‌ها^{۱۲} 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^{۱۳} اشاره کرد.

۲-۲-۱ تاریخچه داده‌کاوی

واژه داده‌کاوی تا اوایل دهه ۹۰ میلادی مفهومی نداشت و بکار برده نمی‌شد. در دهه ۶۰ میلادی متخصصان آمار به جای استفاده از کلمه تحلیل داده، از واژه‌های دیگری مانند صید داده^{۱۴} و لایروبی داده^{۱۵} استفاده می‌کردند. اصل اصطلاح و واژه داده‌کاوی در ابتدا دهه ۹۰ میلادی مورد استفاده قرار گرفت. در دهه ۶۰ میلادی و پیش از آن زمینه‌هایی برای ایجاد سیستم‌های جمع‌آوری و مدیریت داده‌ها ایجاد شد و تحقیقاتی در این زمینه انجام پذیرفت که منجر به معرفی و ایجاد سیستم‌های مدیریت پایگاه داده شد. توسعه سیستم‌های پایگاهی پیشرفته در دهه ۸۰ و ایجاد پایگاه‌های شی‌گرا، کاربردگرا و فعال باعث توسعه همه‌جانبه و کاربردی شدن این سیستم‌ها در سراسر جهان گردید. بدین ترتیب سیستم‌های مدیریت پایگاه داده‌ای همچون Sybase، Oracle، DB2 و غیره ایجاد شدند و حجم زیادی از داده‌ها توسط این سیستم‌ها مورد پردازش قرار گرفت. شاید بتوان مهمترین عامل در معرفی داده‌کاوی را مبحث کشف دانش از پایگاه داده^{۱۶} دانست بطوری که در بسیاری از موارد کشف دانش از پایگاه داده و داده‌کاوی بصورت مترادف بکار برده می‌شوند. الگوریتم‌های داده‌کاوی در دهه اخیر با سرعت بسیار زیاد در حال توسعه هستند.

⁹ Data Mining

¹⁰ Machine Learning

¹¹ Database Management

¹² Data Visualization

¹³ signal processing

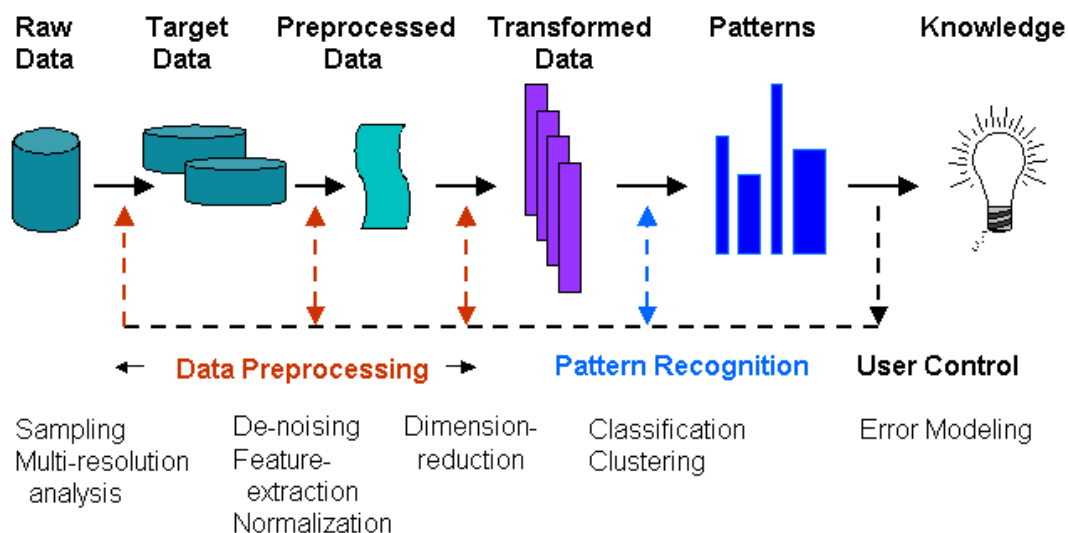
¹⁴ Data Fishing

¹⁵ Data Dredging

¹⁶ Knowledge Discovery From Data

۲-۲-۲ فرآیند داده کاوی

داده کاوی که با عنوان کشف دانش از داده نیز شناخته شده است، به فرایند استخراج اطلاعات و دانش از داده های موجود در پایگاه داده یا انبار داده اطلاق می شود. فرآیند داده کاوی شامل چندین گام است. این فرآیند از داده های خام آغاز می شود و تا شکل دهی دانش جدید ادامه دارد (شکل ۱-۲).



شکل ۱-۲ گام های فرآیند داده کاوی

۳-۲-۲ کاربردهای داده کاوی

داده کاوی تقریباً در تمامی صنایع، کسب و کارها و تجارتها 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 کارها
- ✓ کشف الگوهای رفتاری خرید مشتری در فروشگاه ها و کسب و کارها
- ✓ تحلیل سبد خرید
- ✓ شناسایی مشتریان وفادار
- ✓ آنالیز دقیق نیازهای مشتریان
- ✓ پیش بینی فروش
- ✓ دسته بندی مشتریان بر اساس ملیت، نژاد، زبان، موقعیت مکانی و ...
- ✓ در زمینه های مختلف بانک داری مانند پیش بینی الگوهای 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

۴-۲-۲ یادگیری ماشین

یادگیری ماشین یکی از موضوعات پرکاربرد در زمینه داده کاوی است که زیر مجموعه‌ای از هوش مصنوعی به حساب می‌آید. با استفاده از تکنیک‌های یادگیری ماشین، کامپیوتر، الگوهای موجود در داده‌ها (اطلاعات پردازش شده) را یاد گرفته و می‌تواند از آن استفاده کند.

یادگیری ماشین خود به سه کلاس تقسیم می‌شود: یادگیری نظارت‌یافته^{۱۷}، بدون نظارت^{۱۸} و تقویتی^{۱۹}. در بندهای زیر، این سه روش یادگیری با روش‌های رایج در هر کلاس مورد بحث قرار می‌گیرد و همچنین در شکل ۱-۱ نگاهی کلی به این ۳ کلاس شده است.

۱-۴-۲-۲ 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 عبارتند از: ماشین بردار پشتیبان^{۲۰}، جنگل تصادفی^{۲۱} و شبکه عصبی^{۲۲}.

۲-۴-۲-۲ 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 عبارتند از: خوشه‌بندی K-means^{۲۳} و شبکه‌های عصبی خودسازمان‌یافته^{۲۴}.

۳-۴-۲-۲ 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 عبارتند از: Q learning، DQN^{۲۵} و DDPG^{۲۶} میرجیلی و همکاران [۱].

¹⁷ Supervise Learning

¹⁸ Unsupervised Learning

¹⁹ Reinforcement Learning

²⁰ Support vector machines

²¹ Random forest

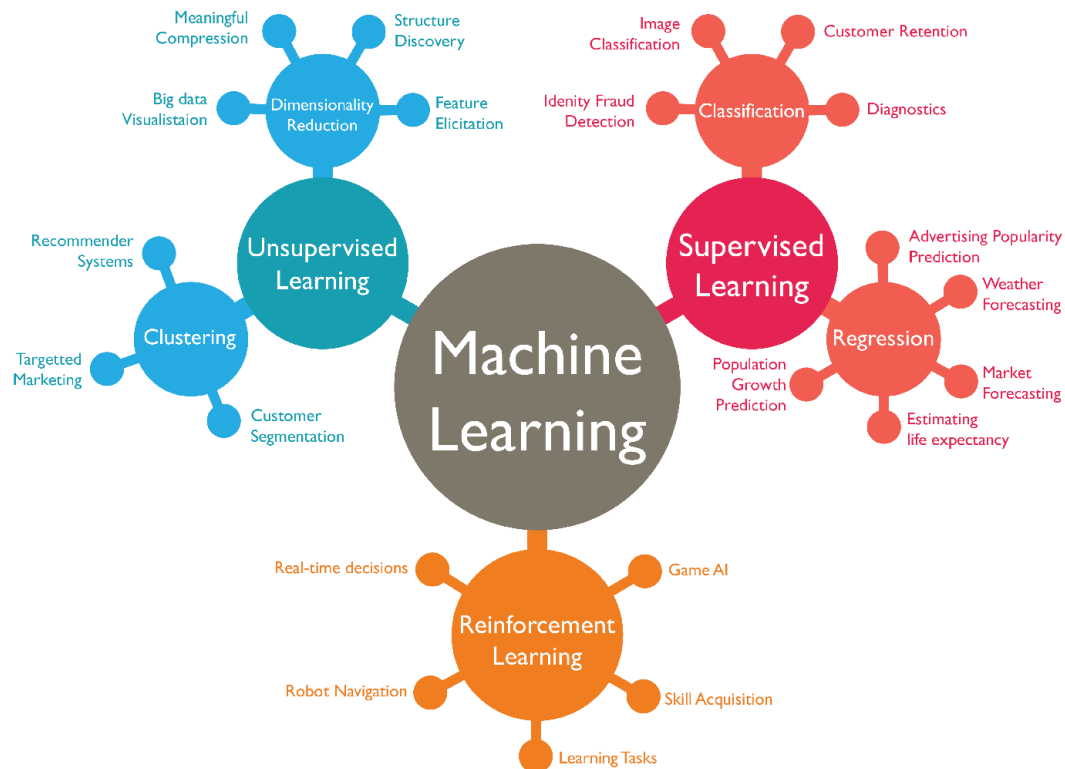
²² neural networks

²³ K-means clustering

²⁴ Self-organizing Neural Networks

²⁵ Deep Q Network

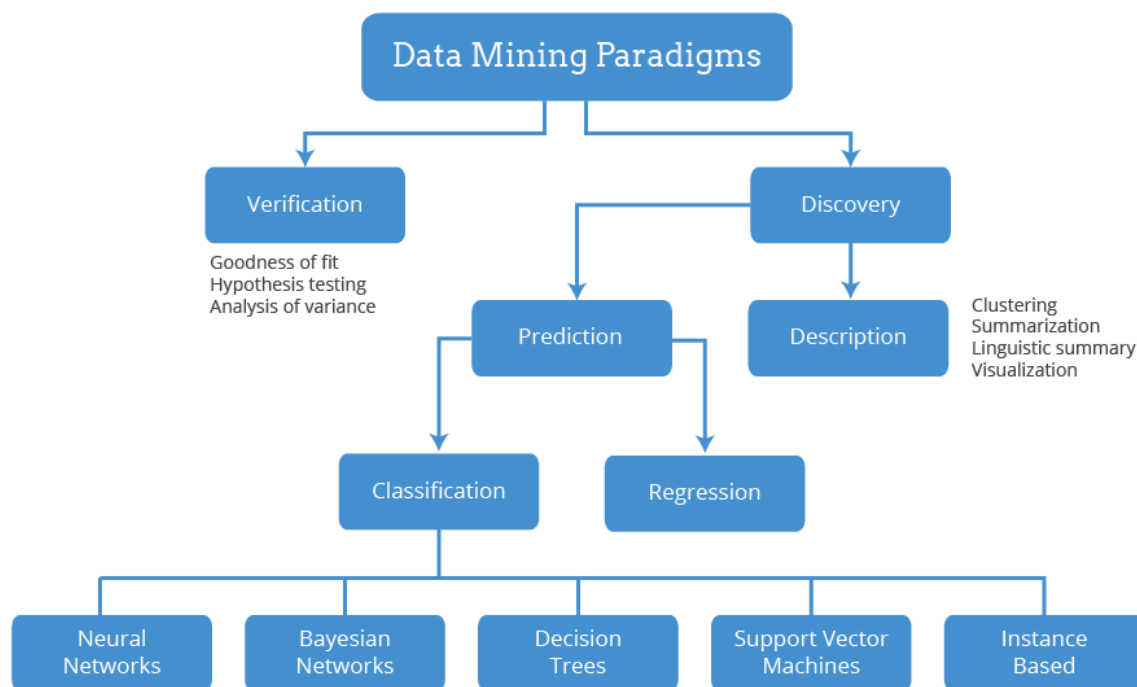
²⁶ Deep Deterministic Policy Gradient



شکل ۲-۲ کلاس‌های مختلف یادگیری ماشین

۳-۲ روش‌های داده‌کاوی

همانطور که در شکل ۲-۲ مشاهده می‌کنید الگوی داده‌کاوی را می‌توان به دو دسته‌ی کلی صحنه‌گذاری و کشف تقسیم کرد. آنچه که امروزه به عنوان روش‌های داده‌کاوی شناخته می‌شود دسته‌ی دوم یعنی کشف می‌باشد. این دسته خود به دو شاخه‌ی کلی پیش‌بینی کننده و توصیف‌کننده تقسیم می‌شود. شاخه‌ی پیش‌بینی کننده خود به دو دسته‌ی طبقه‌بندی و رگرسیون تقسیم می‌شود. از مهم‌ترین روش‌هایی که در شاخه‌ی طبقه‌بندی قرار می‌گیرند می‌توان به شبکه‌های عصبی، شبکه‌های بیزین، درخت تصمیم و ماشین‌های بردار پشتیبان را نام برد. شاخه‌ی توصیف‌کننده نیز خود به شاخه‌هایی همچون خوشه‌بندی، خلاصه‌سازی و مصورسازی تقسیم می‌شود.



شکل ۲-۳ دسته‌بندی روش‌های مختلف داده‌کاوی

۲-۴ تحلیل داده‌های مالی از دیدگاه‌های گوناگون

۲-۴-۱ سیر زمانی تحقیق

نظریه‌های متفاوتی در خصوص ارزیابی و پیش‌بینی بورس در بازارهای 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، تصویری را برای پیش‌بینی قیمت آینده سهام ارائه نمود. مطالعات علمی‌تر با تأکید بر شناسایی دقیق رفتار قیمت سهام، گرایش به سمت مدل‌های ارزشیابی قیمت سهام را به وجود آورد. در ابتدا نظریه‌ی گام‌های تصادفی ۳ 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 کارآی سرمایه شد.

این فرضیه به دلیل ترکیب خاص آن، مورد توجه محافل علمی قرار گرفت. در بازار کارآی سرمایه، اعتقاد بر این است که قیمت سهام انعکاسی از اطلاعات جاری مربوط به آن سهم است و تغییرات قیمت سهام دارای الگوی خاص قابل پیش‌بینی نیست. نظریات مطرح شده تا دهه‌ی ۱۹۸۰ میلادی به‌خوبی تعیین‌کننده‌ی رفتار قیمت سهام در بازار بودند تا اینکه تحولات بازار سهام نیویورک در سال ۱۹۸۷ میلادی، اعتبار فرضیات بازار کارآی سرمایه و مدل‌هایی نظیر تصادفی بودن قیمت‌ها را به‌شدت زیر سؤال برد. در دهه‌ی ۱۹۹۰ میلادی و بعد از آن، بیشتر توجه متخصصان به یک رفتار آشوبگرانه همراه بانظم معطوف شد و تلاش در جهت طراحی مدل‌های غیرخطی به‌منظور پیش‌بینی قیمت سهام اهمیت روزافزونی یافت.

با این نظریات، ازجمله تکنیک‌هایی که اهمیت بالایی یافتند، سیستم‌های هوشمند بودند؛ زیرا با فرض خطی بودن ساختار بازار، به‌آسانی می‌توان بسیاری از مدل‌ها را طراحی نمود. با این وجود، بسیار سخت است که بتوان رفتار مجموعه‌های پیچیده‌ای نظیر بازار سرمایه در یک مجموعه‌ی اقتصادی مدرن را به‌طور کامل در یک مجموعه معادلات ساده و خطی نشان داد. مزیت عمده‌ی سیستم‌های هوشمند نظیر شبکه‌های عصبی مصنوعی و شبکه‌های عصبی فازی، در مدل‌سازی و پیش‌بینی مجموعه‌های نامنظم

و غیرخطی است. ابزار دیگری نظیر الگوریتم ژنتیک نیز از نظر بسیاری از محققان می‌تواند در کاهش زمان به جواب رسیدن و حتی بهینه‌سازی پیش‌بینی‌ها در شبکه‌های عصبی مصنوعی و شبکه‌های عصبی فازی مثر ثمر باشد.

۲-۴-۲ 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 (باش و فخر^[۲]) در سال ۲۰۱۱ 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 می‌شود. 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^{۲۷} استفاده کردند. در پژوهشی دیگر (نصیرطوسی و همکاران^[۳]) 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 ویژگی‌ها با یک روش جدید و ابتکاری، ارائه یک الگوریتم کاهش ویژگی‌ها به نام کاهش ویژگی مبتنی بر هدف^{۲۸} و ارائه یک روش وزن‌دهی احساسات جدید به نام مجموع امتیازات^{۲۹} 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^[۴]) 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 نشانه‌ای 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 نشده‌اند. حال آن‌که استفاده از این داده‌ها می‌تواند به نتایج قابل قبولی منجر شوند. (نصیرطوسی و همکاران^[۵]) در پژوهش خود این موضوع را مورد بررسی قرار دادند و در کار خود، تلاش کرده‌اند تا امکان استفاده از داده‌های بنیادی برای پیش‌بینی حرکت قیمت ارز در بازار فارکس را بررسی کنند. این نوع از پیش‌بینی 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 بنیادی که می‌تواند برای چنین پیش‌بینی‌هایی استفاده شود و یک متدولوژی را پیشنهاد کند، همچنین می‌تواند از طریق آزمایش‌های انجام شده نشان دهد که اگرچه مجموعه‌ای از داده‌های بنیادی ممکن

²⁷ Bagging Trees

²⁸ targeted feature-reduction

²⁹ SumScores

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: پیش پردازش داده‌ها

۳-۱ 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^{۳۰} شوند. پیش‌پردازش نقشی اساسی در روند پردازش داده‌ها و نتایج حاصل از آن‌ها ایفا می‌کند. برای پیش‌پردازش داده‌ها مراحل و ابزارهای مختلفی وجود دارند. برخی از مهم‌ترین مواردی که طی فرایند پیش‌پردازش داده‌ها باید به آن‌ها پرداخته شود عبارتند از:

داده‌های ناموجود^{۳۱}

داده‌های پرت^{۳۲}

نرمال‌سازی داده‌ها^{۳۳}

۳-۲ پیش‌پردازش داده‌ها

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

- ۱) 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.
- ۲) در قسمت بعد ستون تاریخ و ستون متغیر Target2 که شامل رنگ کندل (سبز و قرمز) می‌باشد حذف می‌گردد و به جای آن اعداد ۱ (نماینده رنگ سبز و حرکت صعودی) و ۰ (نماینده رنگ قرمز و حرکت نزولی) قرار داده می‌شوند.
- ۳) در گام بعد (خط ۱۲) از data.describe جهت نرمال استاندارد کردن داده‌ها استفاده شده است.
- ۴) در مرحله بعد به دلیل اینکه داده‌های مربوط به قیمت داده‌های ایستا 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 ۲ 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 ۲۶ ام) از بازدهی لگاریتمی استفاده شده است.
- ۵) 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.
- ۶) برخی از داده‌ها که در هنگام محاسبه لگاریتم مقادیر مثبت یا منفی بی‌نهایت را اختیار می‌کردند در خط ۷۲ حذف شدند.

³⁰ Preprocess

³¹ Missing Data

³² Outliers

³³ Normalization

فصل چهارم: شناسایی ویژگی‌های مورد نظر جهت داده‌کاوی

۱-۴ مقدمه

در مواقعی که بحث کار عملی بر روی داده‌ها پیش می‌آید و از مباحث تئوری فاصله می‌گیریم، شاید مهم‌ترین بخش برای عملیات داده‌کاوی عملیات انتخاب ویژگی^{۳۴} است. در مباحث آکادمیک معمولاً ویژگی‌ها در مسئله در اختیار کاربران قرار دارند ولی در مباحث عملی یک متخصص علوم داده 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 جمع‌آوری 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.

۲-۴ معرفی ویژگی‌های موجود

ویژگی‌های موجود در دیتاست این پژوهش به شرح جدول زیر می‌باشند. همچنین در شکل ۱-۴ می‌توانید اطلاعات مربوط به این ویژگی‌ها که در وبسایت www.kaggle.com موجود است را مشاهده فرمایید.

جدول ۱-۴ ویژگی‌های موجود در مجموعه داده‌ی مورد بررسی 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

³⁴ Feature selection

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

date	1%	Valid ■	75	100%
open	1%	Mismatched ■	0	0%
high	1%	Missing ■	0	0%
low	1%	Unique	75	
Other (71)	95%	Most Common	date	1%

Date (T)	1%	Valid ■	75	100%
Day Open	1%	Mismatched ■	0	0%
Day High	1%	Missing ■	0	0%
Day Low	1%	Unique	75	
Other (71)	95%	Most Common	Date (T)	1%

شکل ۴-۱ اطلاعات ویژگی‌های مورد استفاده در پژوهش

فصل پنجم: انتخاب ویژگی‌های مطلوب

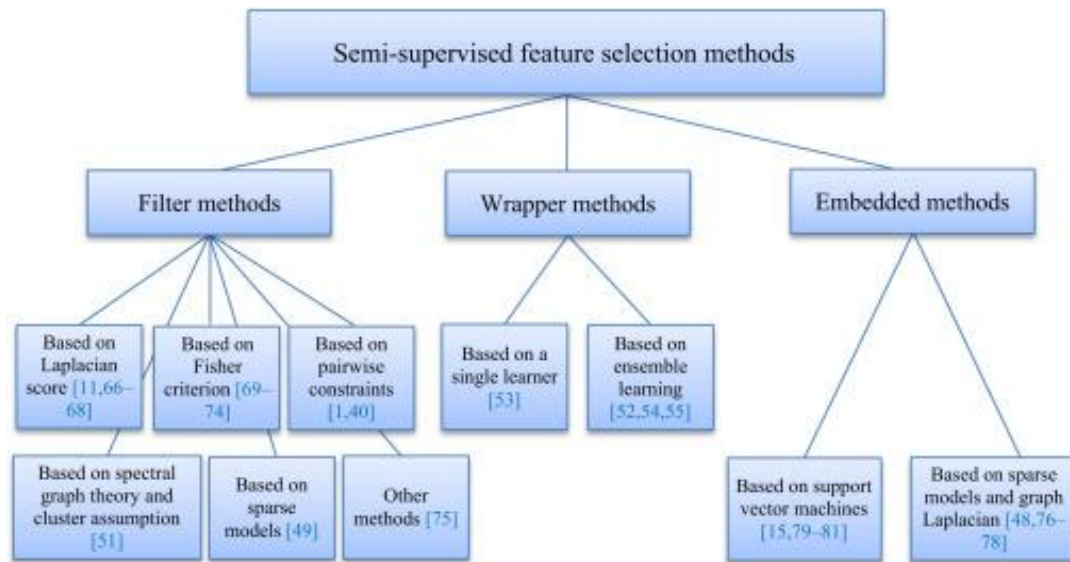
۵-۱ متدولوژی‌های موجود جهت انتخاب ویژگی

روش‌های انتخاب ویژگی^{۳۵} به منظور مواجهه با داده‌های ابعاد بالا، به مولفه‌ای جدایی ناپذیر از فرآیند یادگیری مبدل شده‌اند. به طور کلی همانطور که در شکل ۵-۱ مشخص است روش‌های موجود جهت انتخاب ویژگی‌ها را می‌توان به سه دسته تقسیم کرد که عبارتند از:

«فیلترها»^{۳۶}: این 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»^{۳۷}: 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 نشأت می‌گیرد. با این حال، این تعامل با دسته‌بند منجر به نتایج کارایی بهتری نسبت به فیلترها می‌شود.

«روش‌های توکار»^{۳۸}: انتخاب ویژگی را در فرآیند آموزش انجام می‌دهند و معمولاً 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.



شکل ۵-۱ روش‌های انتخاب ویژگی

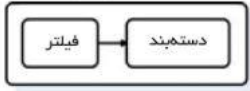

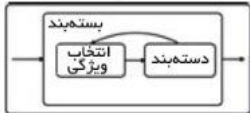
³⁵ Feature Selection Methods

³⁶ Filters

³⁷ Wrappers

³⁸ Embedded

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

معایب	مزایا	روش
فاقد تعامل با دسته‌بندی	استقلال از دسته‌بند هزینه محاسباتی کمتر نسبت به دسته‌بند سریع قابلیت تعمیم خوب	فیلتر 
انتخاب وابسته به دسته‌بندی	تعامل با دسته‌بند هزینه محاسباتی کمتر نسبت به دسته‌بندها ثبت وابستگی ویژگی‌ها	توکار 
به لحاظ محاسباتی گران دارای خطر بیش‌برازش (Overfitting) انتخاب وابسته به دسته‌بند	تعامل با دسته‌بند ثبت وابستگی‌های ویژگی‌ها	بسته‌بند 

شکل ۵-۲ مزایا و معایب روش‌های انتخاب ویژگی

۵-۲ انتخاب ویژگی‌های با بیشترین اثرگذاری

در این قسمت با استفاده از توانایی الگوریتم طبقه‌بندی درخت تصمیم در شناسایی درجه اهمیت پارامترها (توانایی در توضیح نتایج توسط پارامتر مورد نظر) آن‌ها را مرتب می‌کنیم که در نهایت پارامترهایی با درجه اهمیت ۰ را از مدل خارج می‌کنیم.

جدول ۵-۱ مرتب‌سازی ویژگی‌های موجود در مسئله بر اساس میزان اهمیت

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

<i>diff4Strd</i>	0.020225
<i>pctB1Strd</i>	0.0196
<i>rsi2Strd</i>	0.019443
<i>low1_LogDiff</i>	0.019263
<i>close5_LogDiff</i>	0.017952
<i>high2_LogDiff</i>	0.015479
<i>diff0_LogDiff</i>	0.015399
<i>pctB5_LogDiff</i>	0.014492
<i>mavg4_LogDiff</i>	0.014447
<i>up0_LogDiff</i>	0.014187
<i>pctB4_LogDiff</i>	0.013544
<i>pctB4Strd</i>	0.01306
<i>dn1_LogDiff</i>	0.012641
<i>close_LogDiff</i>	0.012304
<i>diff5Strd</i>	0.012183
<i>open3_LogDiff</i>	0.011507
<i>dn0_LogDiff</i>	0.011454
<i>diff5_LogDiff</i>	0.008839
...	...
<i>low2_LogDiff</i>	0
<i>close1_LogDiff</i>	0
<i>high1_LogDiff</i>	0
<i>pctB5Strd</i>	0
<i>open5_LogDiff</i>	0
<i>high5_LogDiff</i>	0
<i>signal0_LogDiff</i>	0
<i>low5_LogDiff</i>	0
<i>diff0Strd</i>	0
<i>mavg0_LogDiff</i>	0
<i>macd1_LogDiff</i>	0
<i>diff1Strd</i>	0
<i>signal2Strd</i>	0
<i>dn3_LogDiff</i>	0
<i>diff2_LogDiff</i>	0
<i>macd3Strd</i>	0
<i>signal3_LogDiff</i>	0
<i>dn5_LogDiff</i>	0
<i>signal4Strd</i>	0
<i>signal4_LogDiff</i>	0
<i>up2_LogDiff</i>	0
<i>diff4_LogDiff</i>	0
<i>macd5Strd</i>	0

<i>macd5_LogDiff</i>	0
<i>signal5Strd</i>	0
<i>signal5_LogDiff</i>	0
<i>dn4_LogDiff</i>	0
<i>mavg5_LogDiff</i>	0
<i>rsi0Strd</i>	0
<i>open_LogDiff</i>	0

فصل ششم: داده‌کاوی و شناسایی
الگوهای پنهان در داده‌ها، و ارائه
نهایی مدل‌های طبقه‌بندی

در این بخش کلیه‌ی مراحل تشریح شده در فصل‌های پیشین و همچنین برخی از مدل‌های پرکاربرد در زمینه طبقه‌بندی بر روی دیتاست معرفی شده در نرم‌افزار پایتون پیاده‌سازی شده است.

```

1 import pandas as pd
2 import numpy as np
3
4
5 data= pd.read_csv("USDJPY_Daily.csv")
6 s=data["target1"][1:]
7 s.reset_index(drop=True,inplace=True)
8 data.drop([len(data)-1],axis=0,inplace=True)
9 data["target1"]=s
10
11 data.drop(["date","target2"],axis=1,inplace=True)
12 data["target1"] = (data["target1"]>0)*1
13
14 ds = data.describe()
15
16 price_list=['open', 'high', 'low', 'close', 'open1', 'high1', 'low1', 'close1',
17            'open2', 'high2', 'low2', 'close2', 'open3', 'high3', 'low3', 'close3',
18            'open4', 'high4', 'low4', 'close4', 'open5', 'high5', 'low5', 'close5',
19            'up5', 'mavg5', 'mavg4', 'dn4', 'up4', 'mavg2', 'up2', 'dn5',
20            'dn3', 'mavg3', 'up3', 'dn0', 'mavg0', 'up0', 'dn2', 'up1']
21
22 new_data=pd.DataFrame()
23
24
25 #calculating logarithmic return of price data time series
26 for i in price_list:
27     s=np.log(data[i][1:])
28     s.reset_index(drop=True, inplace=True)
29     m = np.log(data[i][:-1])
30     m.reset_index(drop=True, inplace=True)
31     new_data[i+" _LogDiff"] = s-m
32     new_data[i]= data[i][1:]
33
34
35 other_features_list=['macd0', 'signal0', 'diff0', 'macd1', 'signal1', 'diff1', 'macd2',
36                    'signal2', 'diff2', 'macd3', 'signal3', 'diff3', 'macd4', 'signal4',
37                    'diff4', 'macd5', 'signal5', 'diff5', 'rsi0', 'rsi1', 'rsi2', 'rsi3',
38                    'rsi4', 'rsi5', 'pct80', 'dn1', 'mavg1',
39                    'pct81', 'pct82', 'pct83',
40                    'pct84', 'pct85']
41
42 for i in other_features_list:
43     s=(data[i][1:]-ds[i]["mean"])/(ds[i]["std"])
44     s.reset_index(drop=True, inplace=True)
45     new_data[i+"Strd"]=s
46     s = np.log(data[i][1:])
47     s.reset_index(drop=True, inplace=True)
48     m = np.log(data[i][:-1])
49     m.reset_index(drop=True, inplace=True)
50     new_data[i+" _LogDiff"]=s-m
51     s=data[i][1:]
52     s.reset_index(drop=True, inplace=True)
53     new_data[i]=s
54
55 s=data["target1"][1:]
56 s.reset_index(drop=True, inplace=True)
57 new_data["target1"]=s
58
59 #drop unimportant features

```

```

59 #drop unimportant features
60 new_data.drop(price_list,axis=1,inplace=True)
61 new_data.drop(other_features_list,axis=1,inplace=True)
62 new_data.drop(["macd2Strd","macd0Strd","signal0Strd","macd1Strd","signal1Strd","signal1_LogDiff",
63               "diff1_LogDiff","up1_LogDiff","signal2_LogDiff","macd3_LogDiff","signal3Strd",
64               "diff3Strd","macd4Strd","macd4_LogDiff"],axis=1,inplace=True)
65
66
67
68 s=new_data["high_LogDiff"][1:]
69 s.reset_index(drop=True,inplace=True)
70 new_data.drop([len(new_data)-1],axis=0,inplace=True)
71 new_data.reset_index(drop=True,inplace=True)
72 new_data["high_LogDiff"]=s
73
74 new_data.drop([0],axis=0,inplace=True)
75 new_data.reset_index(drop=True,inplace=True)
76
77
78 new_data=new_data.dropna()
79 new_data.reset_index(drop=True,inplace=True)
80 print("data len is :",len(new_data))
81 x_train,y_train = new_data.iloc[:int(len(new_data)*0.8):-1],new_data.iloc[:int(len(new_data)*0.8):-1]
82 x_text, y_text = new_data.iloc[int(len(new_data)*0.8):-1],new_data.iloc[int(len(new_data)*0.8):-1]
83
84 from sklearn.linear_model import LogisticRegression
85 from sklearn.neighbors import KNeighborsClassifier
86 from sklearn.tree import DecisionTreeClassifier
87 from sklearn.svm import SVC
88 from sklearn.ensemble import RandomForestClassifier
89
90 LR_classifier = LogisticRegression(random_state=0)
91 KNC_classifier = KNeighborsClassifier(n_neighbors=5)
92 DTC_classifier=DecisionTreeClassifier()
93 SVC_classifier = SVC(kernel="linear", random_state=0)
94 RFC_classifier = RandomForestClassifier(n_estimators=10,criterion='entropy',random_state=0)
95
96 LR_classifier.fit(x_train,y_train)
97 KNC_classifier.fit(x_train,y_train)
98 DTC_classifier.fit(x_train,y_train)
99 SVC_classifier.fit(x_train,y_train)
100 RFC_classifier.fit(x_train,y_train)
101
102 LR_p=LR_classifier.predict(x_text)
103 KNC_p=KNC_classifier.predict(x_text)
104 DTC_p=DTC_classifier.predict(x_text)
105 SVC_p=SVC_classifier.predict(x_text)
106 RFC_p=RFC_classifier.predict(x_text)
107
108
109
110
111 featureImportance = pd.DataFrame(DTC_classifier.feature_importances_,index=x_train.columns,columns=["FI"])
112 featureImportance.sort_values(["FI"],ascending=False,inplace=True)
113 print(featureImportance)
114
115
116 from sklearn.metrics import confusion_matrix
117
118
119 LR_cm= confusion_matrix(y_text,LR_p)
120 print("Logistic Regression Confusion Matrix:")
121 print(LR_cm)
122 print((LR_cm[0][0]+LR_cm[1][1])/(LR_cm[0][0]+LR_cm[1][1]+LR_cm[0][1]+LR_cm[1][0]))
123
124 KNC_cm= confusion_matrix(y_text,KNC_p)
125 print("KNeighbors Classifier Confusion Matrix:")
126 print(KNC_cm)
127 print((KNC_cm[0][0]+KNC_cm[1][1])/(KNC_cm[0][0]+KNC_cm[1][1]+KNC_cm[0][1]+KNC_cm[1][0]))
128
129 DTC_cm= confusion_matrix(y_text,DTC_p)
130 print("DecisionTree Classifier Confusion Matrix:")
131 print(DTC_cm)
132 print((DTC_cm[0][0]+DTC_cm[1][1])/(DTC_cm[0][0]+DTC_cm[1][1]+DTC_cm[0][1]+DTC_cm[1][0]))
133
134
135 SVC_cm= confusion_matrix(y_text,SVC_p)
136 print("SVC Classifier Confusion Matrix:")
137 print(SVC_cm)
138 print((SVC_cm[0][0]+SVC_cm[1][1])/(SVC_cm[0][0]+SVC_cm[1][1]+SVC_cm[0][1]+SVC_cm[1][0]))
139
140
141
142
143 RFC_cm= confusion_matrix(y_text,RFC_p)
144 print("RandomForest Classifier Confusion Matrix:")
145 print(RFC_cm)
146 print((RFC_cm[0][0]+RFC_cm[1][1])/(RFC_cm[0][0]+RFC_cm[1][1]+RFC_cm[0][1]+RFC_cm[1][0]))
147
148

```

شکل ۷-۱ پیاده‌سازی مراحل تشریح شده در نرم افزار پایتون

```

Python 3.7.4 (default, Aug 9 2019, 18:34:13) [MSC v.1915 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 7.8.0 -- An enhanced Interactive Python.

In [1]: runfile('C:/Users/AmerAndish/Downloads/4_5868646130752948302.py', wdir='C:/Users/AmerAndish/Downloads')
C:\Users\AmerAndish\Anaconda3\lib\site-packages\pandas\core\series.py:853: RuntimeWarning: invalid value encountered in log
  result = getatrr(ufunc, method)(*inputs, **kwargs)
data len is : 598
C:\Users\AmerAndish\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py:432: FutureWarning: Default solver will be
changed to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)

      FI
high_LogDiff      0.280191
pctB0_LogDiff     0.077747
pctB2Strd         0.043798
open1_LogDiff     0.039968
rsi0_LogDiff      0.033280
...
signal3_LogDiff   0.000000
signal4Strd       0.000000
signal4_LogDiff   0.000000
diff4_LogDiff     0.000000
pctB5_LogDiff     0.000000

[90 rows x 1 columns]
Logistic Regression Confusion Matrix:
[[22 40]
 [18 40]]
0.5166666666666667
KNeighbors Classifier Confusion Matrix:
[[29 33]
 [25 33]]
0.5166666666666667
DecisionTree Classifier Confusion Matrix:
[[35 27]
 [19 39]]
0.6166666666666667
SVC Classifier Confusion Matrix:
[[21 41]
 [19 39]]
0.5
RandomForest Classifier Confusion Matrix:
[[37 25]
 [23 35]]
0.6

In [2]:

```

شکل ۷-۲ نتایج حاصل از مدل

نتایج حاصل از مدل‌های اعمال شده بوسیله ماتریس درهم‌ریختگی^{۳۹} نمایش داده می‌شود. این ماتریس که در شکل ۷-۱ نشان داده شده است از ۴ قسمت تشکیل شده است که عبارتند از:

- بخش شمال غربی: داده‌های مثبتی که به درستی مثبت پیش‌بینی شده‌اند. که در این تحقیق شامل روزهایی است که روند حرکت جفت‌ارز صعودی بوده و و مدل نیز به درستی صعودی پیش‌بینی کرده است.
- بخش شمال شرقی: داده‌های منفی که به اشتباه مثبت پیش‌بینی شده‌اند. که در این تحقیق شامل روزهایی است که روند حرکت جفت‌ارز نزولی بوده و و مدل به اشتباه صعودی پیش‌بینی کرده است.
- بخش جنوب غربی: داده‌های مثبتی که به اشتباه منفی پیش‌بینی شده‌اند. که در این تحقیق شامل روزهایی است که روند حرکت جفت‌ارز صعودی بوده و و مدل به اشتباه نزولی پیش‌بینی کرده است.

³⁹ Confusion matrix

- بخش جنوب شرقی: داده‌های منفی که به درستی منفی پیش‌بینی شده‌اند. که در این تحقیق شامل روزهایی است که روند حرکت جفت‌ارز نزولی بوده و و مدل نیز به درستی نزولی پیش‌بینی کرده است.

		Condition Phase (Worst Case)		
		Condition Positive/ Shaded	Condition Negative/ Unshaded	
Testing Phase (Best Case)	Test Positive/ Shaded	True positive shaded T_p (Correct)	False positive shaded F_p (Incorrect)	Precision/Positive Predictive Value (PPV) $\frac{T_p}{T_p + F_p} \times 100\%$
	Test Negative/ Unshaded	False negative unshaded F_n (Incorrect)	True negative unshaded T_n (Correct)	Negative Predictive Value (NPV) $\frac{T_n}{T_n + F_n} \times 100$
		Sensitivity/Recall Rate (RR) $\frac{T_p}{T_p + F_n} \times 100\%$	Specificity Rate (SR) $\frac{T_n}{T_n + F_p} \times 100\%$	

شکل ۷-۳ ماتریس درهم‌ریختگی

همچنین با استفاده از مقادیر موجود در این ماتریس می‌توان مقادیر شاخص‌های صحت^{۴۰}، حساسیت^{۴۱}، وضوح^{۴۲} و در نهایت میزان دقت^{۴۳} مدل را محاسبه کرد.

بنابر نتایج حاصل شده پس از پیاده‌سازی مدل میزان دقت مدل‌ها بر اساس ماتریس درهم‌ریختگی برابر است با:

$$\text{Logistic Regression Confusion Matrix} = \begin{bmatrix} 22 & 40 \\ 18 & 40 \end{bmatrix} \rightarrow \text{Accuracy} = \frac{22 + 40}{22 + 40 + 18 + 40} = 0.516$$

$$\text{KNeighbors Classifier Confusion Matrix} = \begin{bmatrix} 29 & 33 \\ 25 & 33 \end{bmatrix} \rightarrow \text{Accuracy} = \frac{29 + 33}{29 + 33 + 25 + 33} = 0.516$$

⁴⁰ precision

⁴¹ sensitivity

⁴² specificity

⁴³ Accuracy

$$\begin{aligned} \text{Decision Tree Classifier Confusion Matrix} &= \begin{bmatrix} 35 & 27 \\ 19 & 39 \end{bmatrix} \rightarrow \text{Accuracy} \\ &= \frac{35 + 27}{35 + 27 + 19 + 39} = 0.616 \end{aligned}$$

$$\begin{aligned} \text{SVC Classifier Confusion Matrix} &= \begin{bmatrix} 21 & 41 \\ 19 & 39 \end{bmatrix} \rightarrow \text{Accuracy} = \frac{21 + 41}{21 + 41 + 19 + 39} \\ &= 0.50 \end{aligned}$$

$$\begin{aligned} \text{RandomForest Classifier Confusion Matrix} &= \begin{bmatrix} 37 & 25 \\ 23 & 35 \end{bmatrix} \rightarrow \text{Accuracy} \\ &= \frac{37 + 25}{37 + 25 + 23 + 35} = 0.60 \end{aligned}$$

فصل هفتم: نتیجه‌گیری

در این پژوهش به بررسی بازار جهانی فارکس و پیش‌بینی روند حرکت یکی از جفت‌ارزهای این بازار یعنی دلار-ین پرداخته شده است. در گام اول به بررسی کلی داده‌های موجود بوسیله نرم‌افزار R پرداختیم تا یک دید کلی از وضعیت داده‌ها حاصل شود. در ادامه جهت دستیابی به نتیجه‌ی بهتر از مدل به آماده‌سازی داده‌ها پرداخته شد. پس از آن از بین پارامترهای موجود برخی از پارامترها با درجه اهمیت بالاتر برای استفاده در مدل‌های طبقه‌بندی گزینش شدند و در گام آخر نیز برخی از مدل‌های پرکاربرد در طبقه‌بندی شامل رگرسیون لجستیک^{۴۴}، درخت تصادفی، ماشین بردار پشتیبان، جنگل تصادفی و k همسایه^{۴۵} برای طبقه‌بندی داده‌ها و پیش‌بینی روند حرکت جفت-ارز مذکور در روز آینده مود بررسی قرار گرفتند که نتیجه‌ی حاصل از روش‌های درخت تصمیم و جنگل تصادفی از بالاترین دقت برخوردار بودند. البته لازم به ذکر است که بدلیل ماهیت داده‌های موجود و عدم ایستایی در داده‌ها پیش‌بینی آن‌ها بسیار دشوار است و در اکثر مقالات مورد بررسی نیز دقت مدل‌های حاصل از حد مشخصی (حدود ۷۰٪) تجاوز نکرده است که این موضوع خود حاکی از دشواری پیش‌بینی در چنین فضایی است و همین امر دلیل خوبی برای نیاز به تحقیقات بیشتر و گسترده تر در این زمینه می‌باشد.

⁴⁴ Logistic regression

⁴⁵ K neighbors

پیوست: کد نرم افزار و توضیح درباره نرم افزار مورد استفاده

❖ نرم افزار مورد استفاده جهت پیاده سازی مدل



پایتون یک زبان برنامه نویسی قدرتمند سطح بالا، شی 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، واسطه‌های برنامه‌نویسی و... به کار گرفته می‌شود. این زبان با قابلیت‌های فراوان و شگفت‌انگیزی که دارد تحولی در دنیای برنامه نویسی از توسعه برنامه‌های تحت وب تا ایجاد بازی‌های رایانه‌ای، بوجود آورده است. پایتون ابتدا در سال ۱۹۹۱ 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 تصور کرد 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 کاملاً 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 کاملاً قابل درک است. این زبان این 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 کاربردهای این زبان محبوب می‌توان به موارد زیر اشاره کرد:

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

❖ کد نرم افزار

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

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. Fakh. *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.