فهرست مطالب

[مقدمه 3](#_Toc51538634)

[داده کاوی چیست؟ 3](#_Toc51538635)

[تاریخچه داده‌کاوی 4](#_Toc51538636)

[چ**را داده‌کاوی؟** 4](#_Toc51538637)

[پیش‌بینی خودکار گرایش‌ها و رفتارها 4](#_Toc51538638)

[کشف خودکار الگوهای پیش‌تر ناشناخته 5](#_Toc51538639)

[انواع منابع داده 5](#_Toc51538640)

[تراکنش‌های کسب‌و‌کار 5](#_Toc51538641)

[داده‌های علمی 5](#_Toc51538642)

[داده‌های شخصی و پزشکی 6](#_Toc51538643)

[تصاویر و ویدئوهای نظارتی 6](#_Toc51538644)

[رقابت‌های ورزشی 6](#_Toc51538645)

[رسانه‌های دیجیتال 6](#_Toc51538646)

[دنیاهای مجازی 7](#_Toc51538647)

[جهان‌های مجازی 7](#_Toc51538648)

[گزارش‌ها و اسناد متنی 7](#_Toc51538649)

[فرایند داده‌کاوی 7](#_Toc51538650)

[پاک‌سازی داده 8](#_Toc51538651)

[یکپارچه‌سازی داده‌ها 9](#_Toc51538652)

[انتخاب داده 9](#_Toc51538653)

[تبدیل داده 9](#_Toc51538654)

[داده‌کاوی 9](#_Toc51538655)

[ارزیابی الگو 9](#_Toc51538656)

[ارائه دانش 9](#_Toc51538657)

[مشکلات داده‌کاوی 10](#_Toc51538658)

[مسائل روش‌شناسی داده‌کاوی 10](#_Toc51538659)

[مسائل کارایی 10](#_Toc51538660)

[مسائل منابع داده 10](#_Toc51538661)

[مزایا و معایب داده‌کاوی 11](#_Toc51538662)

[مزایای داده‌کاوی 11](#_Toc51538663)

[معایب داده‌کاوی 12](#_Toc51538664)

[تاثیرات مثبت 12](#_Toc51538665)

[**اثرات منفی** 12](#_Toc51538666)

[کاربردهای داده‌کاوی 12](#_Toc51538667)

[ده الگوریتم برتر داده کاوی 13](#_Toc51538668)

[الگوریتم C4.5 13](#_Toc51538669)

[الگوریتم k-means 14](#_Toc51538670)

[مشکلات روش خوشه‌ بندی K-Means 14](#_Toc51538671)

[الگوریتم Support vector machines 14](#_Toc51538672)

[الگوریتم Apriori 15](#_Toc51538673)

[الگوریتم EM 15](#_Toc51538674)

[الگوریتم PageRank 15](#_Toc51538675)

[الگوریتم AdaBoost 16](#_Toc51538676)

[الگوریتم kNN 16](#_Toc51538677)

[الگوریتم Naive Bayes 16](#_Toc51538678)

[الگوریتم CART 17](#_Toc51538679)

[داده‌ها 17](#_Toc51538680)

[آماده‌سازی داده: 17](#_Toc51538681)

[الگوریتم‌های مورد استفاده 19](#_Toc51538682)

[الگوریتم Lasso 19](#_Toc51538683)

[الگوریتم Ridge 20](#_Toc51538684)

[الگوریتم LinearSVR 21](#_Toc51538685)

[الگوریتم Decision Tree 22](#_Toc51538686)

[ارزیابی الگوریتم‌ها 23](#_Toc51538687)

# مقدمه

# داده کاوی چیست؟

به مجموعه‌ای از روش‌های قابل اعمال بر پایگاه داده‌های بزرگ و پیچیده به منظور کشف الگوهای پنهان و جالب توجه نهفته در میان داده‌ها، داده‌کاوی گفته می‌شود. روش‌های داده‌کاوی تقریبا همیشه به لحاظ محاسباتی پر هزینه هستند. علم میان‌رشته‌ای داده‌کاوی، پیرامون ابزارها، متدولوژی‌ها و تئوری‌هایی است که برای آشکارسازی الگوهای موجود در داده‌ها مورد استفاده قرار می‌گیرند و گامی اساسی در راستای کشف دانش محسوب می‌شود. دلایل گوناگونی پیرامون چرایی مبدل شدن داده‌کاوی به چنین حوزه مهمی از مطالعات وجود دارد. برخی از این موارد در ادامه بیان شده‌اند.

۱. رشد انفجاری داده‌ها در گستره وسیعی از زمینه‌ها در صنعت و دانشگاه که توسط موارد زیر پشتیبانی می‌شود:

* دستگاه‌های ذخیره‌سازی نسبت به گذشته ارزان‌تر و با ظرفیت نامحدود، مانند فضاهای ذخیره‌سازی ابری
* ارتباطات سریع‌تر با سرعت اتصال بیشتر
* سیستم‌های مدیریت پایگاه داده و پشتیبانی نرم‌افزاری بهتر

۲. قدرت پردازش کامپیوتری به سرعت در حال افزایش:

با چنین حجم بالا و متنوعی از داده‌های موجود، روش‌های داده‌کاوی به استخراج اطلاعات از داده‌ها کمک می‌کنند.

روش‌های داده‌کاوی دارای انواع گوناگونی هستند و از [**رگرسیون**](https://blog.faradars.org/tag/%D8%B1%DA%AF%D8%B1%D8%B3%DB%8C%D9%88%D9%86/) گرفته تا روش‌های تشخیص الگوی پیچیده و دارای هزینه محاسباتی بالا که ریشه در علوم کامپیوتر دارند را شامل می‌شوند. هدف اصلی روش‌های یادگیری (داده‌کاوی) انجام پیش‌بینی است، ولی این تنها هدف داده‌کاوی نیست.

# تاریخچه داده‌کاوی

در سال ۱۹۶۰، کارشناسان آمار از اصطلاحات «صید داده» (Data Fishing) و «لایروبی داده» (Data Dredging) برای ارجاع به فعالیت‌های «تحلیل داده» (Data Analytics) استفاده می‌کردند. اصطلاح «داده‌کاوی» در حدود سال ۱۹۹۰ در جامعه پایگاه‌داده مورد استفاده قرار گرفت و به محبوبیت قابل توجهی دست پیدا کرد. عنوان مناسب‌تر برای فرآیند داده‌کاوی، «کشف دانش از داده» (Knowledge Discovery From Data) است.

# چ**را داده‌کاوی؟**

با رشد و افزایش توجهات به داده‌کاوی، پرسش «چرا داده‌کاوی؟» همواره مطرح می‌شود. در پاسخ به این پرسش باید گفت، داده‌کاوی دارای کاربردهای زیادی است. بدین ترتیب، زمینه‌ای جوان و آینده‌دار برای نسل کنونی محسوب می‌شود. این زمینه توانسته توجهات زیادی را به صنایع و جوامع اطلاعاتی جلب کند. با وجود گستره وسیع داده‌ها، نیاز حتمی به تبدیل چنین داده‌هایی به اطلاعات و دانش وجود دارد.

بنابراین، بشر از اطلاعات و دانش برای گستره وسیعی از کاربردها، از تحلیل بازار گرفته تا تشخیص بیماری‌ها، کشف کلاهبرداری و پیش‌بینی قیمت سهام استفاده می‌کند. در مجموع باید گفت، ضر‌ب‌المثل انگلیسی «نیاز، مادر همه ابداعات بشر است»، پاسخی کوتاه و گویا به پرسش مطرح شده است. در ادامه، برخی از استفاده‌های داده‌کاوی مورد بررسی قرار گرفته‌اند.

## پیش‌بینی خودکار گرایش‌ها و رفتارها

از داده‌کاوی برای خودکارسازی فرآیندها و انجام پیش‌بینی در پایگاه‌داده‌های بزرگ استفاده می‌شود. پرسش‌هایی که پاسخگویی به آن‌ها نیازمند تحلیل‌های گسترده است، اکنون و با استفاده از تحلیل داده‌ها قابل پاسخگویی هستند. بازاریابی هدفمند مثالی از بازاریابی پیش‌بین است. همچنین، از داده‌کاوی برای ارسال ایمیل‌های تبلیغاتی هدفمند و بهینه استفاده می‌شود. در واقع، داده‌کاوی به منظور بیشینه‌سازی «بازگشت سرمایه» (Return On Investment) در ارسال ایمیل‌های تبلیغاتی مورد استفاده قرار می‌گیرد. از دیگر مسائل پیش‌بینی می‌توان به پیش‌بینی ورشکستگی، اشاره کرد. شناسایی بخش‌هایی از جامعه که احتمال دارد به یک رویداد واکنش‌های مشابهی نشان دهند نیز از دیگر قابلیت‌های داده‌کاوی به شمار می‌آید.

## کشف خودکار الگوهای پیش‌تر ناشناخته

از ابزارهای داده‌کاوی برای بررسی پایگاه‌های داده استفاده می‌شود. همچنین، برای شناسایی الگوهای از پیش ناشناخته نیز قابل بهره‌برداری است. یک مثال خیلی خوب از کاوش الگوها، تحلیل داده‌های فروش خرده‌فروشی‌ها است. این کار با هدف شناسایی محصولات غیر مرتبطی که معمولا با هم خریداری می‌شوند انجام می‌شود. همچنین، مسائل کاوش الگوی دیگری نیز وجود دارند که از جمله آن‌ها می‌توان به شناسایی تراکنش‌های کلاهبرداری در کارت‌های اعتباری اشاره کرد. در چنین مواردی، الگوهای داده ناشناخته و جدید، می‌توانند خبر از وقوع سرقت اطلاعات کارت اعتباری و دیگر انواع کلاهبرداری بدهند.

# انواع منابع داده

در این بخش، انواع منابع داده‌ای که همه روزه حجم انبوهی از داده‌ها را تولید و یا ذخیره می‌کنند، مورد بررسی قرار گرفته‌اند. در همین راستا، مطالعه مطلب [**«مجموعه داده‌های رایگان و قابل دانلود برای علم داده و یادگیری ماشین»**](https://blog.faradars.org/open-datasets-for-machine-learning-and-data-science/) به علاقمندان توصیه می‌شود.

## تراکنش‌های کسب‌و‌کار

در کسب‌و‌کارهای کنونی، اغلب تراکنش‌ها تا ابد نگهداری می‌شوند. بسیاری از این تراکنش‌ها دارای زمان هستند و شامل معاملات درون کسب‌و‌کاری مانند خریدها، مبادلات بانکداری، سهام و دیگر موارد هستند.

## داده‌های علمی

در سراسر جهان، جوامع گوناگون در حال گردآوری حجم انبوهی از داده‌های علمی هستند. این داده‌های علمی نیاز به تحلیل دارند. این در حالی است که همواره نیاز به ثبت داده‌های جدید بیشتر با سرعت بالاتری وجود دارد. داده‌کاوی در زمینه‌های علمی گوناگون برای کمک به تحلیل داده‌ها و کشف دانش از آن‌ها کمک شایان توجهی می‌کند.

## داده‌های شخصی و پزشکی

داده‌ها، از شخصی گرفته تا عمومی و از فردی گرفته تا دولتی را می‌توان با اهداف گوناگونی گردآوری کرد و مورد تحلیل قرار داد. این داده‌ها برای افراد و گروه‌های مختلف مورد نیاز هستند و هنگامی که گردآوری شدند، کشف اطلاعات از آن‌ها می‌تواند پرده از مسائل مهمی بردارد. از جمله داده‌های شخصی، می‌توان به اطلاعات تراکنش‌های بانکی فرد و یا اسناد پزشکی ایشان اشاره کرد. داده‌کاوی در داده‌های پزشکی نقش قابل توجهی در پیشگیری، کشف و حتی درمان بیماری‌ها دارد.

## تصاویر و ویدئوهای نظارتی

با کاهش قیمت دوربین‌های عکاسی و فیلمبرداری و وجود دوربین در گوشی‌های هوشمند، در هر لحظه حجم زیادی از داده‌های چندرسانه‌ای تولید می‌شود. از سوی دیگر، حجم زیادی از تصاویر و ویدئوها نیز توسط دوربین‌های نظارتی گردآوری می‌شوند. این داده‌ها برای انواع تحلیل‌های داده قابل بهره‌برداری هستند.

## رقابت‌های ورزشی

حجم زیادی از داده‌ها و آمارها پیرامون رقابت‌های ورزشی وجود دارد که قابل گردآوری و تحلیل محسوب می‌شوند. از جمله این موارد می‌توان به اطلاعات بازی و بازیکنان اشاره کرد.

## رسانه‌های دیجیتال

دلایل زیادی منجر به انفجار مخازن داده‌های دیجیتال شده است. از جمله این موارد می‌توان به اسکنرهای ارزان، دوربین‌های ویدئویی دسکتاپ و دوربین‌های دیجیتال اشاره کرد. از سوی دیگر، شرکت‌های بزرگی مانند NHL و NBA در حال حاضر کار تبدیل مجموعه‌های خود به داده‌های دیجیتال را آغاز کرده‌اند و انجام چنین کارهایی نیز نیاز به تحلیل حجم انبوه داده‌ها را برجسته‌تر می‌سازد.

## دنیاهای مجازی

سیستم‌های «طراحی به کمک کامپیوتر» (Computer Aided Design) متعددی برای معماران وجود دارند. این سیستم‌ها برای تولید حجم انبوهی از داده‌ها مورد استفاده قرار می‌گیرند. علاوه بر این، می‌توان از داده‌های «مهندسی نرم‌افزار» (Software Engineering) به عنوان منبعی از داده‌ها – همراه با کدهای فراوان برای امور گوناگون – استفاده کرد.

## جهان‌های مجازی

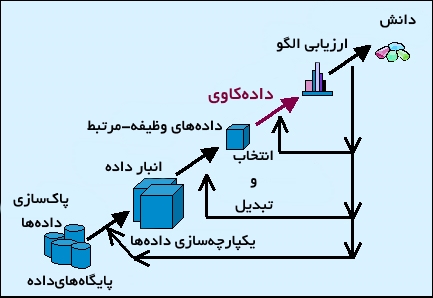
امروزه بسیاری از برنامه‌های کاربردی از فضاهای مجازی سه‌بُعدی استفاده می‌کنند. همچنین، این فضاها و اشیایی که در بر می‌گیرند باید با زبان خاصی مانند «زبان مدل‌سازی واقعیت مجازی» (Virtual Reality Modeling Language | VRML) توصیف شوند.

## گزارش‌ها و اسناد متنی

ارتباطات در بسیاری از شرکت‌ها بر مبنای گزارش‌ها و اسناد دارای قالب متنی است. این اسناد برای انجام تحلیل‌های آتی قابل نگهداری هستند. از سوی دیگر، حجم انبوهی از داده‌های موجود در وب نیز به صورت داده‌های متنی ساختار نیافته هستند که هر روز بر حجم آن‌ها افزوده می‌شود.

# فرایند داده‌کاوی

داده‌کاوی که با عنوان «کشف دانش از داده» (Knowledge Discovery From Data | KDD) نیز شناخته شده است، فرایند استخراج اطلاعات و دانش از داده‌های موجود در پایگاه‌داده یا انبارداده است.



فرآیند داده‌کاوی شامل چندین گام است. این فرآیند از داده‌های خام آغاز می‌شود و تا شکل‌دهی دانش جدید ادامه دارد. فرآیند بازگشتی داده‌کاوی شامل گام‌های زیر است:

* «پاک‌سازی داده» (Data Cleaning)
* «یکپارچه‌سازی داده» (Data Integration)
* «انتخاب داده» (Data Selection)
* «تبدیل داده» (Data Transformation)
* «کاوش داده» (Data Mining)
* «ارزیابی الگو» (Pattern Evaluation)
* «ارائه دانش» (Knowledge Representation)

## پاک‌سازی داده

در این فاز «نویز» (نوفه) از مجموعه حذف و تدابیری برای «داده‌های ناموجود» (Missing Values) اندیشیده می‌شود.

## یکپارچه‌سازی داده‌ها

در اغلب مسائل داده‌کاوی، داده‌ها از منابع داده گوناگون باید به یکباره مورد تحلیل قرار بگیرند. مثال خوبی از این مورد پایگاه داده‌های شعب مختلف یک فروشگاه زنجیره‌ای در شهرها و کشورهای گوناگون جهان است. برای تحلیل این داده‌ها باید آن‌ها را به صورت یکپارچه در یک [**«انبار داده» (Data Warehouse)**](https://blog.faradars.org/data-warehouse-2/) گردآوری کرد، این کار در فاز یکپارچه‌سازی انجام می‌شود.

## انتخاب داده

در فاز انتخاب داده، باید داده‌های مرتبط با تحلیل انتخاب، و از مجموعه داده برای انجام تحلیل‌ها بازیابی شوند.

## تبدیل داده

تبدیل داده یک روش تثبیت داده نیز هست. در این فاز، داده‌های انتخاب شده به فرم دیگری تبدیل می‌شوند. این کار به ساده‌تر شدن، بهبود صحت و دقت فرآیند کاوش کمک می‌کند.

## داده‌کاوی

در این فاز از روش‌های هوشمندانه برای استخراج الگوهای مهم و جالب توجه از میان داده‌ها استفاده می‌شود.

## ارزیابی الگو

در این فرآیند، الگوهای حاصل شده در گام قبل، از جنبه‌های گوناگونی شامل دقت، صحت و قابلیت تعمیم و دیگر موارد مورد ارزیابی قرار می‌گیرند.

## ارائه دانش

ارائه دانش فاز نهایی فرآیند داده‌کاوی است. در این فاز، دانش کشف شده به شیوه قابل درک به کاربر ارائه می‌شود. در این گام حیاتی و بسیار مهم، روش‌های بصری‌سازی مورد استفاده قرار می‌گیرند. این کار به کاربران در درک و تفسیر نتایج داده‌کاوی کمک می‌کند.

# مشکلات داده‌کاوی

در این بخش از راهنمای داده‌کاوی، برخی از مسائل کلی که داده‌کاوی با آن‌ها مواجه است تشریح شده‌اند.

## مسائل روش‌شناسی داده‌کاوی

این چالش به روش‌های موجود برای داده‌کاوی و محدودیت‌های آن‌ها مانند تطبیق‌پذیری مربوط است. در واقع، ارائه روش‌هایی که دارای پیچیدگی کم و قابلیت تعمیم به مسائل گوناگون باشند و در عین حال بتوانند با حجم انبوهی از داده‌ها کار کنند از جمله مسائل مربوط به بحث روش‌شناسی در داده‌کاوی است.

## مسائل کارایی

روش‌های هوش مصنوعی و آماری زیادی وجود دارند که در داده‌کاوی مورد استفاده قرار می‌گیرند. اغلب این روش‌ها برای مجموعه داده‌های خیلی بزرگ طراحی نشده‌اند و این چالشی است که داده‌کاوی این روزها با آن دست و پنجه نرم می‌کند. زیرا امروزه حجم داده‌ها از ترابایت، پتابایت و اگزابایت نیز عبور کرده است. می‌توان گفت این امر موجب افزایش مسائل مرتبط با مقیاس‌پذیری و کارایی روش‌های داده‌کاوی می‌شود و نیاز به روش‌هایی را ایجاد می‌کند که بتوانند به طور قابل توجهی داده‌های بزرگ را پردازش کنند.

در چنین شرایطی، ممکن است از نمونه‌برداری به جای کل مجموعه‌داده استفاده شود. اگرچه، در این حالت نیز مسائلی مانند کامل بودن و روش انتخاب نمونه‌ها بروز می‌کند. دیگر موضوع در بحث کارایی به‌روزرسانی تدریجی و برنامه‌نویسی موازی است. از موازی‌سازی برای حل مساله اندازه استفاده می‌شود و طی آن اگر مجموعه داده به زیرمجموعه‌هایی تقسیم شود، نتایج بعدا قابل ادغام شدن هستند. به‌روز‌رسانی مداوم برای ادغام نتایج از «کاوش موازی» (Parallel Mining) بسیار حائز اهمیت است. داده‌های جدید بدون نیاز به بازتحلیل کل مجموعه داده در دسترس قرار می‌گیرند.

## مسائل منابع داده

مسائل زیادی در رابطه با منابع داده لازم/مورد استفاده برای داده‌کاوی وجود دارد. برخی از این مسائل مانند تنوع داده‌ها کاربردی و بخشی دیگر مسائل فلسفی‌تری مانند مشکل انباشته شدن داده‌ها هستند. واضح است که در حال حاضر حجم زیادی از داده‌ها، بیش از آنکه قابل مدیریت باشند، وجود دارند. از سوی دیگر، بشر همچنان در حال گردآوری داده‌ها حتی با نرخ بالاتری است. گسترش سیستم‌های مدیریت پایگاه داده یکی از عواملی بوده که به رشد گردآوری داده‌ها کمک شایان توجهی کرده است.

ظهور داده‌کاوی قطعا منجر به برداشت داده‌های بیشتری می‌شود. با توجه به مسائل کاربردی مرتبط با منابع داده، پایگاه داده‌های موضوعی ایجاد شده‌اند. بدین ترتیب، نیاز به تمرکز کردن روی انواع داده پیچیده‌تر وجود دارد. انواع گوناگونی از داده‌ها در گستره متنوعی از مخازن ذخیره‌سازی می‌شوند. سخت است که بشر انتظار داشته باشد یک سیستم داده‌کاوی نتایج کاوش خوبی را برای همه داده‌ها و منابع داده کسب کند.

داده‌ها و منابع گوناگون داده ممکن است نیاز به الگوریتم‌ها و متدولوژی‌های متمایزی داشته باشند. در حال حاضر، تمرکز بر پایگاه داده‌های رابطه‌ای و انبارهای داده است. ابزارهای داده‌کاوی نیز طیف گسترده‌ای را برای انواع داده‌ها شامل می‌شوند. علاوه بر این، منابع داده، در سطح ساختاری و معنایی، چالش‌های مهمی را به همراه دارند. این تنها به جامعه پایگاه‌داده مربوط نیست، بلکه به جامعه داده‌کاوی نیز ارتباط دارد.

# مزایا و معایب داده‌کاوی

داده‌کاوی دارای مزایا و معایب متعددی است. برخی از این موارد در ادامه بیان شده‌اند.

## مزایای داده‌کاوی

* برای پیدا کردن کلاهبرداری‌های احتمالی از داده‌کاوی در بانک‌ها و موسسات مالی استفاده می‌شود. این کار بر پایه تراکنش‌ها، رفتار کاربر و الگوهای داده‌ها انجام می‌شود.
* به تبلیغ‌کنندگان کمک می‌کند تا تبلیغات درستی را در اینترنت قرار دهند. این کار در صفحات وب و برپایه الگوریتم‌های یادگیری ماشین انجام می‌شود. بدین شکل داده‌کاوی هم به خریداران و هم فروشندگان محصولات و خدمات سود می‌رساند.
* فروشگاه‌های خرده‌فروشی و مواد غذایی از داده‌کاوی برای چینش قفسه‌های فروشگاهی و تحلیل سبد خرید مشتریان خود استفاده می‌کنند. بدین شکل داده‌کاوی به افزایش درآمد آن‌ها کمک می‌کند.
* از داده‌کاوی در زمینه‌های گوناگونی از جمله «بیوانفورماتیک» (bio-informatic)، پزشکی و ژنتیک با اهداف گوناگون شامل پیشگیری، تشخیص و درمان بیماری‌ها استفاده می‌شود.
* داده‌کاوی توسط سازمان‌های قانونی برای شناسایی مظنون‌های جنایی مورد استفاده قرار می‌گیرند.

## معایب داده‌کاوی

* روش‌های داده‌کاوی ۱۰۰٪ صحیح نیستند. بنابراین ممکن است در برخی شرایط عواقب بسیار بدی را در پی داشته باشند.
* کار با برخی سیستم‌ها و روش‌های داده‌کاوی دشوار و نیازمند دانش قابل توجهی است.
* برخی از مسائل داده‌کاوی حریم خصوصی و حتی امنیت کاربر را تحت تاثیر قرار می‌دهند.

در مجموع می‌توان تاثیرات مثبت و منفی زیر را برای داده‌کاوی بیان کرد.

### **تاثیرات مثبت**

* پیش‌بینی گرایش‌های آینده
* کمک در تصمیم‌گیری
* بهبود درآمد سازمان‌ها و کاهش هزینه‌های آن‌ها
* تحلیل سبد خرید
* تشخیص کلاهبرداری

### **اثرات منفی**

* حریم خصوصی/امنیت کاربر
* حجم غافلگیرکننده داده‌ها
* هزینه بالا در گام پیاده‌سازی
* استفاده‌های احتمالی از اطلاعات
* عدم صحت احتمالی داده‌ها

# کاربردهای داده‌کاوی

از آنجا که داده‌ها اغلب بسیار ارزان و روش‌های گردآوری داده تقریبا به طور کامل خودکارسازی شده‌اند، در بسیاری از زمینه‌ها، مانند کسب‌و‌کار، موفقیت بستگی به استفاده موثر و هوشمندانه از داده‌های گردآوری شده دارد. در همین راستا باید گفت که تلاش ها در حوزه داده‌کاوی در زمینه‌های گوناگونی در حال وقوع است. مثال‌هایی که در ادامه می‌آید، تنها نشانگر برخی از حوزه‌های کاربرد جالب داده‌کاوی است. هر چه ارتباطات بیشتری میان رشته‌های گوناگون به وقوع بپیوندد، دامنه کاربردها تکامل یافته و کاربردهای جدیدی ظهور می‌کنند. برخی از کاربردهای داده‌کاوی در ادامه بیان و برای چندی از آن‌ها توضیحاتی ارائه شده است.

* تجارت الکترونیک
* خودروهای خودران
* خطرات درمان‌های جدید
* پژوهش‌های فضایی
* تشخیص کلاهبرداری
* تحلیل تجارت سهام
* پیش‌بینی کسب‌و‌کار
* شبکه‌های اجتماعی
* تحلیل مشتریان

# ده الگوریتم برتر داده کاوی

الگوریتم C4.5   
الگوریتم C4.5 یک دسته‌بندی (classifier) را در قالب یک درخت تصمیم تولید می‌کند که دارای ۲ نوع گره است. یک گره به‌صورت برگ که یک دسته را مشخص می‌کند و یک گره تصمیم که آزمون‌هایی روی یک صفت انجام می‌دهد تا یک شاخه یا زیر درخت به ازای هر خروجی آزمون تولید می‌کند. حالا

classifier چیست؟  
واژه classifier مفهومی فراتر و کلی تر از کلاس را دارد که علاوه بر کلاس، واسط ها و انواع داده ای را نیز پوشش می دهد.

الگوریتم C4.5 بهینه شده الگوریتم ID3 می باشد که از قانون هرس بعدی بهره می برد و می تواند صفاتی را که داده های نویزی و مقدار و همچنین صفات گسسته ندارند، استفاده نماید. در C4.5 فرض بر این است که کل داده های آموزشی در داخل حافظه باشند.  
به جهت ساخت درخت تصمیم، فرض می کنیم که مجموعه داده های آموزشی که دارای برچسب کلاس مربوطه و بردار ویژگی ها هستند، در دسترس می باشند. معیارهای گوناگونی برای تقسیم بندی گره ها در درخت تصمیم وجود دارد که از عمومی ترین آنها، معیار ضریب بهره اطلاعات است که در C4.5 به کار می رد.  
درخت تصمیم بر پایه آنالیز داده های ورودی و برای یافتن یک ویژگی بر مبنای تصمیم گیری برای هر نود استفاده می شود. ویژگی های گوناگونی از داده در هر نود بررسی می شود و یک ویژگی که اگر انتخاب شود، باعث خواهد شد که بی نظمی (آنتروپی) کاهش یابد، گزینش می شود. مبنای فعالیت نیز بر این اساس ایجاد شده است.

الگوریتم k-means   
این روش علی‌رغم سادگی آن‌یک روش پایه برای بسیاری از روش‌های خوشه‌بندی دیگر محسوب می‌شود. این روش روشی انحصاری و مسطح محسوب می‌شود. برای این الگوریتم شکل‌های مختلفی بیان‌شده است. ولی همه آن‌ها دارای روالی تکراری هستند که برای تعدادی ثابت از خوشه‌ها سعی در تخمین موارد زیر دارند:  
به دست آوردن نقاطی به‌عنوان مراکز خوشه‌ها این نقاط درواقع همان میانگین نقاط متعلق به هر خوشه هستند.  
نسبت دادن هر نمونه داده به یک خوشه که آن داده کمترین فاصله تا مرکز آن خوشه را دارا باشد.

مشکلات روش خوشه‌ بندی K-Means  
علی‌رغم اینکه خاتمه پذیری الگوریتم بالا تضمین‌شده است ولی جواب نهایی آن واحد نبوده و همواره جوابی بهینه نمی‌باشد. به‌طورکلی روش ساده بالا دارای مشکلات زیر است.  
جواب نهایی به انتخاب خوشه‌های اولیه وابستگی دارد.  
روالی مشخص برای محاسبه اولیه مراکز خوشه‌ها وجود ندارد.  
اگر در تکراری از الگوریتم تعداد داده‌های متعلق به خوشه‌ای صفر شد راهی برای تغییر و بهبود ادامه روش وجود ندارد. در این روش فرض شده است که تعداد خوشه‌ها از ابتدا مشخص است. اما معمولاً در کاربردهای زیادی تعداد خوشه‌ها مشخص نمی‌باشد.

الگوریتم Support vector machines   
ماشین بردار پشتیبانی (Support vector machines – SVMs) یکی از روش‌های یادگیری بانظارت است که از آن برای طبقه‌بندی و رگرسیون استفاده می‌کنند.  
این روش ازجمله روش‌های نسبتاً جدیدی است که در سال‌های اخیر کارایی خوبی نسبت به روش‌های قدیمی‌تر برای طبقه‌بندی ازجمله شبکه‌های عصبی پرسپترون نشان داده است. مبنای کاریدسته‌بندی کننده SVM دسته‌بندی خطی داده‌ها است و در تقسیم خطی داده‌ها سعی می‌کنیم خطی را انتخاب کنیم که حاشیه اطمینان بیشتری داشته باشد. حل معادل پیدا کردن خط بهینه برای داده‌ها به‌وسیله روش‌های QP که روش‌های شناخته‌شده‌ای در حل مسائل محدودیت دار هستند صورت می‌گیرد. قبل از تقسیمِ خطی برای اینکه ماشین بتواند داده‌های با پیچیدگی بالا را دسته‌بندی کند داده‌ها را به‌وسیلهٔ تابعِ phi به فضای با ابعاد خیلی بالاتر می‌بریم. برای اینکه بتوانیم مسئله ابعاد خیلی بالا را با استفاده از این روش‌ها حل کنیم از قضیه دوگانی لاگرانژ برای تبدیلِ مسئلهٔ مینیمم‌سازی موردنظر به فرم دوگانی آن‌که در آن به‌جای تابع پیچیدهٔ phi که ما را به فضایی با ابعاد بالا می‌برد، تابعِ ساده‌تری به نامِ تابع هسته که ضرب برداری تابع phi است ظاهر می‌شود استفاده می‌کنیم. از توابع هسته مختلفی ازجمله هسته‌های نمایی، چندجمله‌ای و سیگموید می‌توان استفاده نمود.

الگوریتم Apriori  
آپریوری یک الگوریتم کلاسیک برای یادگیری قوانین وابستگی است. آپریوری روی پایگاه‌های داده شامل تراکنش‌ها (مثلاً مجموعه محصولات خریداری شده توسط مشتریان در یک سوپرمارکت) ساخته شده‌است. الگوریتم‌های دیگری نیز در این زمینه وجود دارند که روی پایگاه داده‌هایی کار می‌کنند که یا شامل تراکنش نیستند (Winepi و Minepi) و یا دارای ثبت زمانی نیستند (DNA sequencing).  
ورودی این الگوریتم مجموعه‌ای از مجموعه آیتم‌ها است. الگوریتم تلاش می‌کند تا زیرمجموعه‌هایی از آیتم‌ها را که حداقل بین C مجموعه آیتم مشترک است بیابد. آپریوری یک الگوریتم پایین به بالا است، آنگونه که در هر مرحله یک آیتم به زیرمجموعه‌های مکرر اضافه می‌شود. مجموعه کاندیدها روی داده مورد ارزیابی قرار می‌گیرند. شرط خاتمه الگوریتم، عدم وجود شیوه توسعه موفق دیگری است.  
هدف الگوریتم آپریوری، یافتن وابستگی‌ها بین مجموعه‌های مختلف از داده‌است. گاهی به آن، تحلیل سبد خرید هم می‌گویند. هر مجموعه‌ای از داده تعدادی آیتم دارد و تراکنش نامیده می‌شود.

الگوریتم EM  
الگوریتم امید ریاضی-بیشینه‌سازی (EM) یک روش تکرارشونده (iterative) است که به دنبال یافتن برآوردی با بیشترین درست نمایی برای پارامترهای یک توزیع پارامتری است. این الگوریتم روش متداول برای زمان‌هایی است که برخی از متغیرهای تصادفی پنهان هستند.

الگوریتم PageRank  
پیج‌رنک یا رتبه صفحه (به انگلیسی: PageRank) به فناوری گفته می‌شود که بر پایه آن موتورهای جستجویی همچون گوگل وب‌گاه‌هایی که به هدف جستجوگر نزدیک‌ترند را در رده‌های بالاتری نسبت به دیگران قرار می‌دهد. اصطلاح انگلیسی این مفهوم به دنبال لری پیج نویسنده اولیه این الگوریتم و از مؤسسین اولیه گوگل نام‌گذاری شده است. به این روش کاربرانی که کلمه ویژه‌ای را جستجو می‌کنند می‌توانند ابتدا وب‌گاه‌هایی را ببینند که هم به خواسته آن‌ها نزدیک‌تر است و هم بازدید بیشتری داشته است.

پیج رنک که به‌اختصار PR می‌گویند برای اولین بار در سال ۱۹۹۶ میلادی توسط Sergey Brin ابداع گردید، پیج رنک نتیجه تجزیه‌وتحلیل لینک‌های ورودی به یک سایت است، پیج رنک رتبه‌ای است که گوگل برای یک سایت بین اعداد ۰ تا ۱۰ در نظر می‌گیرد. هرقدر این عدد نزدیک به ۱۰ باشد نشان‌دهنده این است که این سایت لینک‌های ورودی زیادی دارد و سایتهای زیادی به این سایت لینک داده‌اند. روش دیگر موتورهای جستجو، پردازش رتبه صفحه، با استفاده از تعداد یافته‌های خود در سایتهای اینترنتی می‌باشد. این روش، یک فناوری زیرساخت برای پدیدار گشتن «بمب گوگلی» نیز بود. ناگفته نماند که هدف از ساختن «بمب‌های گوگلی» بالاتر رفتن رتبه یک وبلاگ یا وب‌گاه در موتور جستجوی گوگل می‌باشد که معمولاً با همکاری گروه‌های مردمی ایجاد می‌شود.

البته به‌جز گوگل شرکت‌های دیگری هم به رتبه دهی به سایتها می‌پردازند، که یکی از آن‌ها الکسا (به انگلیسی: Alexa) (یک سایت معتبر بین‌المللی و دارای نمودار رتبه جهانی برای سایت‌ها می‌باشد) نام دارد که الگوریتم بالا بردن رتبه سایت‌ها را بر اساس تعداد درست بازدیدکنندگان انجام می‌دهد. این رتبه‌بندی شماره‌ای بین ۱ تا چند میلیون می‌باشد که هرچه این شماره به یک نزدیک‌تر باشد رتبه وب‌سایت شما بالاتر می‌باشد.

الگوریتم AdaBoost  
آدابوست مخفف بوستینگ تطبیقی بوده و یک الگوریتم یادگیری ماشین است که توسط یاو فروند و رابرت شاپیر ابداع شد. درواقع آدابوست یک متا الگوریتم است که بمظور ارتقاء عملکرد، و رفع مشکل رده‌های نامتوازن همراه دیگر الگوریتم‌های یادگیری استفاده می‌شود. در این الگوریتم، طبقه بند هر مرحله جدید به نفع نمونه‌های غلط طبقه‌بندی‌شده در مراحل قبل تنظیم می‌گردد.

الگوریتم kNN  
جستجوی نزدیک‌ترین همسایه یا Nearest Neighbor، که همچنین با نامه‌ای جستجوی مجاورت، جستجوی همسانی یا جستجوی نزدیک‌ترین نقطه شناخته می‌شود، یک مسئله بهینه‌سازی برای پیدا کردن نزدیک‌ترین نقطه‌ها در فضاهای متریک است. مسئله بدین‌صورت است که: مجموعه S شامل تعدادی نقطه در یک فضای متریک مانند M و نیز یک نقطهٔ پرس و جوی q∈ M داده‌شده است، هدف پیدا کردن نزدیک‌ترین نقطه در S به q است. در بسیاری از موارد، فضای M به‌صورت یک فضای اقلیدسی d-بعدی و فاصله بین نقاط با معیار فاصله اقلیدسی، فاصله منهتن یا دیگر فاصله‌های متریک سنجیده می‌شود.  
جستجوی k نزدیک‌ترین همسایه، K همسایه نزدیک‌تر به نقطه پرس‌وجو را برمی‌گرداند. این روش معمولاً در تجزیه‌وتحلیل پیش‌بینی، به‌منظور تخمین و یا دسته‌بندی یک نقطه بر اساس اجماع همسایگان آن استفاده می‌شود. گراف k نزدیک‌ترین همسایه گرافیست که در آن هر نقطه در گراف K نزدیک‌ترین همسایگان خود متصل است.

الگوریتم Naive Bayes  
به طور ساده روش بیز روشی برای دسته بندی پدیده‌ها، بر پایه احتمال وقوع یا عدم وقوع یک پدیده‌است.  
بر اساس ویژگی‌های ذاتی احتمال (به‌ویژه اشتراک احتمال) نایو بیز (به انگلیسی: Naive Bayes classifier) با دریافت تمرین اولیه نتایج خوبی ارائه خواهد کرد. شیوه یادگیری در روش نایو بیز از نوع یادگیری با ناظر (به انگلیسی: Supervised learning) است.  
برای نمونه یک میوه ممکن است پرتغال باشد. اگر نارنجی و کروی با شعاع حدود ده سانتی‌متر باشد. اگر این احتمالات به‌درستی به همدیگر وابسته باشند نایو بیز در تشخیص اینکه این میوه پرتغال است یا نه به‌درستی عمل خواهد کرد.  
برنامه‌های کاربردی بسیاری هستند که پارامترهای نایو بیز را تخمین می‌زنند، بنابراین افراد بدون سروکار داشتن با تئوری بیز می‌توانند از این امکان به‌منظور حل مسائل موردنظر بهره ببرند. باوجود مسائل طراحی و پیش‌فرض‌هایی که در خصوص روش بیز وجود دارد، این روش برای طبقه‌بندی کردن بیشتر مسائل در جهان واقعی، مناسب است.

الگوریتم CART  
این روش که موجب تشکیل یک درخت تصمیم با تقسیمات دوتایی می‌گردد، توسط بریمن و همکارانش در سال ۱۹۸۴ به‌طور کامل معرفی شد. این روش برای متغیرهای کمی طراحی گردیده ولی قابل‌استفاده برای هر نوع متغیری است. بر اساس این الگوریتم، نرم‌افزار آماری تحت نام CART نیز ساخته‌شده است که از شناخته‌شده‌ترین برنامه‌ها است. در این روش و برای متغیر پاسخ کیفی، شاخص جینی ()Gini Index به‌عنوان معیاری برای انتخاب متغیرهای مناسب، معرفی‌شده است.

# داده‌ها

داده‌ها‌ی مورد استفاده در این مسئله شامل گزارشات بیماران، فوتی‌ها و افراد بهبودیافته از بیماری کوید۱۹ هستند که از کشورهای سراسر جهان در طی ۲۲۲ روز جمع‌اوری شده‌است (به عبارت دقیق‌تر از ۲۲ ژانویه تا ۲۹ آگوست). از آنجا که ۱) تعداد کشورهای درگیر در ابتدا و در انتهای تاریخ متفاوت است، ۲) همچنین نحوه‌ی گزارش دهی کشورهای مختلف در تعداد بیماران و تعداد بهبود یافتگان متفاوت است و ۳) به دلیل ماهیت نهفته‌ی بیماری در درصد بالاتری از افراد، لذا هدف این تحقیق یافتن و مطالعه‌ی روند تعداد مرگ و میر در این بازه‌ی زمانی‌است.

## آماده‌سازی داده:

برای داده‌ی موجود نیاز بود که با استفاده از pivot تعداد مرگ و میر کشور‌های و مناطق مختلف در روزها تجمیع شود. با استفاده از pivot تعداد ۹۸ هزار ردیف به ۲۲۲ ردیف داده تقلیل یافت و به شکل مشابه برای ماه‌های سال. در شکل و نمودار زیر روند و تعداد مرگ و میر مشاهده می‌شود.

Figure 1نمودارتعداد مرگ‌ومیر در طی ۲۲۲ روز

در شکل ۱ نمودار تعداد مرگ‌ومیر از ۲۲ ژانویه تا آخر آگوست به تفکیک روز مشاهده می‌شود. همچنین جدول ۱ تعداد مرگ و میر در هر ماه را نشان می‌دهد.

Table 1نمودار تعداد مرگ و میر به تفکیک ماه

|  |  |  |  |
| --- | --- | --- | --- |
| Months | Sum of Confirmed | Sum of Deaths | Sum of Recovered |
| Jan | 38535 | 889 | 844 |
| Feb | 1671959 | 46898 | 380794 |
| Mar | 8894726 | 386173 | 2707659 |
| Apr | 63402405 | 4176098 | 16414806 |
| May | 144769885 | 9568332 | 54214094 |
| Jun | 246578791 | 13285645 | 118668091 |
| Jul | 431490739 | 18390651 | 242205298 |
| Aug | 620429831 | 22185051 | 390019904 |

# الگوریتم‌های مورد استفاده

برای انجام این تحقیق از روش‌های Lasso Regression، Ridge Regression، LinearSVR و Decision Tree Regression استفاده شده است. تمامی الگوریتم‌ها با استفاده از python 3.6.8 و در محیط برنامه نویسی jupyter پیاده‌سازی و ارزیابی شده‌اند.

چون داده‌های مورد استفاده مطابق شکل ۱ اکیدا صعودی بوده و شبیه به نمودار نمایی است. از طرفی الگوریتم‌های مذکور روی داده‌ی نرمال عملکرد بهتری دارند پس لازم است شکل داده‌ی ورودی مقدار تغییر کند تا به شکل نرمال نزدیک‌تر باشد. لذت در این تحقیق، لگاریتم طبیعی () تعداد مرگ و میر روزانه به عنوان هدف برازش الگوریتم‌های مذکور درنظر گرفته‌شده‌است. بدین ترتیب بعد از اعمال الگوریتم، برای مقایسه و رسم نمودارها‌ی روند پیش بینی، خروجی الگوریتم‌‌ها با معکوس لگاریتم (یعنی عدد طبیعی به توان خروجی‌ها) محاسبه شده و نمایش داده‌شده است.

هر کدام از الگوریتم‌های بالا دارای ابر پارامترهایی هستند که جداگانه تنظیم شده و مناسبترین پارامتر استفاده شده‌است.

## الگوریتم Lasso

این روش یک الگوریتم درون‌یابی است که دارای عبارت تنظیمی با درجه‌ی یک می‌باشد که ضریب این عبارت یک ابرپارامتر مهم جهت تنظیم و استفاده از آن است. در این روش بعد از تنظیم پارامترها مطابق شکل ۱ مقدار بهینه‌ی Alpha = 10 بدست آمد و برای این مقدار پیش‌بینی به صورت شکل ۲ می‌باشد که در آن خط آبی سیر مرگ و میر در ۲۰۰ روز اول و خط نارنجی در مابقی روز‌هاست. پیش‌بینی‌ این الگوریتم

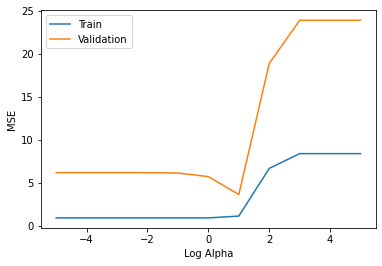


Figure 2 یافتن مقدار بهینه‌ی alpha

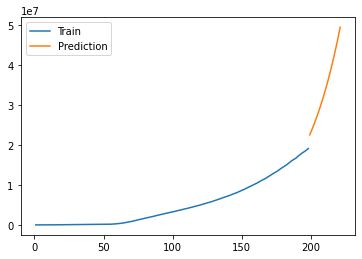


Figure 3 روند پیش بینی Lasso

## الگوریتم Ridge

این روش یک الگوریتم درون‌یابی است که دارای عبارت تنظیمی با درجه‌ی دو می‌باشد که ضریب این عبارت یک ابرپارامتر مهم در زمان استفاده از آن است. در این روش بعد از تنظیم پارامترها همانند شکل ۴ مقدار بهینه‌ی Alpha = 1000 بدست آمد و برای این مقدار پیش‌بینی به صورت شکل ۵ می‌باشد که در آن خط آبی سیر مرگ و میر در ۲۰۰ روز اول و خط نارنجی در مابقی روز‌هاست.

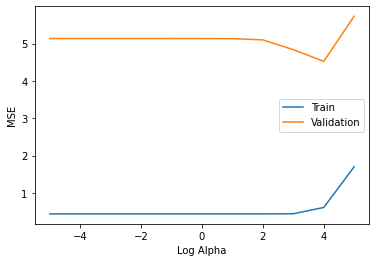


Figure 4 یافتن مقدار بهینه‌ی alpha

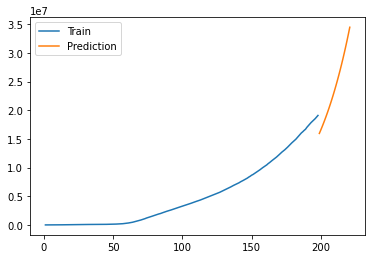


Figure 5 روند پیش بینی Ridge

## الگوریتم LinearSVR

در این روش بعد از تنظیم پارامتر مقدار هزینه مطابق شکل ۶ مقدار بهینه‌ی C = 10 بدست آمد و برای این مقدار پیش‌بینی به صورت شکل ۷ می‌باشد که در آن خط آبی سیر مرگ و میر در ۲۰۰ روز اول و خط نارنجی در مابقی روز‌هاست.

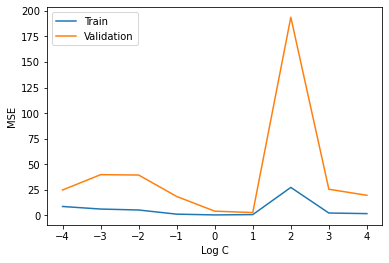


Figure 6 یافتن مقدار بهینه‌ی C

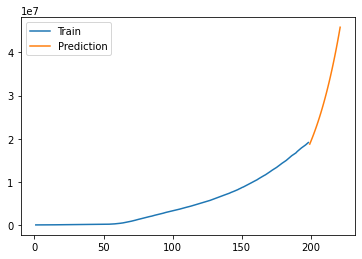


Figure 7 روند پیش بینی LinearSVR

## الگوریتم Decision Tree

در این روش بعد از تنظیم پارامترها مقدار بهینه‌ی max\_depth = 10 و min\_samples\_split =100 بدست آمد و برای این مقدار پیش‌بینی به صورت شکل ۸ می‌باشد که در آن خط آبی سیر مرگ و میر در ۲۰۰ روز اول و خط نارنجی در مابقی روز‌هاست.

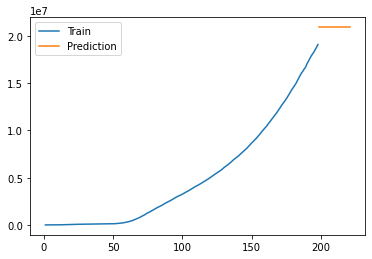


Figure 8 روند پیش‌بینی Decision Tree

# ارزیابی الگوریتم‌ها

در جدول ۲ نتایج الگوریتام‌ها در قالب یک جدول آماده است. این الگوریتم‌ها از منظر MSE بر روی داده‌ی اعتبارسنجی با هم مقایسه‌شده‌اند. این عبارت مقدار میانگین خطای‌ نقاط پیش‌بینی برروی داده‌های اعتبارسنجی می‌باشد.

Table 2 جدول مقایسه‌ی نتایج براساس مقادیر خطای MSE و RMSE. در این جدول هر چه این مقادیر کمتر باشند الگوریتم خواهد بود

|  |  |
| --- | --- |
| MSE |  |
| 2.656 | Lasso Regression |
| 4.524 | Ridge Regression |
| 1.517 | Linear Support Vector Regression (SVR) |
| 0.672 | Decision Tree Regression |

براساس جدول ۲ الگوریتم درخت تصمیم خطای کمتری دارد، اما با توجه به این‌که مقدار پیش‌بینی شده تقریبا ثابت است، جواب‌های این الگوریتم نمی‌تواند پیش بینی مناسبی برای این داده ارائه دهد. لذا الگوریتم LinearSVR بنظر نتایج بهتری از دو منظر پیش‌بینی‌های مصور شده از روند تعداد مرگ و میر طبق شکل ۷ ارائه کرده‌است و از دیدگاه MSE و خطای مدل نیز مقادیر قابل قبولی را ارائه می‌دهد. قدرت این الگوریتم از پشتوانه‌ی ریاضی قدرتمند آن یعنی بردارهای پشتیبان (Support Vector) نشات می‌گیرد.