**فهرست**

**مقدمه و پیشینه ای در زمینه داده کاوی3**

**ضرورت و اهمیت داده کاوی4**

**کاربرد های داده کاوی و موارد استفاده از آن5**

**مزایا و معایب داده کاوی10**

**مراحل و استراتژی های موجود در داده کاوی13**

**ابزار ها و روش های موجود در داده کاوی19**

فهرست شکل ها

شکل 1 .............................................................................................15

شکل 2 .............................................................................................17

شکل 3 .............................................................................................22

شکل 4 .............................................................................................24

شکل 5 .............................................................................................25

مقدمه و پیشینه ای در زمینه داده کاوی

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.

داده کاوی در واقع یک روش است، یک روش که قرار است به واسطه آن مسئله‌ای حل شود. این روش بر روی حجم زیادی از داده‌ها کار ‌می‌کند و تحلیل مورد نیاز را انجام ‌می‌دهد، در نهایت یک سری الگوهای تکرارشونده استخراج ‌می‌شود که ‌می‌تواند برگ برنده باشد. سپس نوبت به پیدا کردن ارتباطات بین الگوهای مختلف است و در نهایت یک سری راه حل‌های مهم در این چالش وارد ‌می‌شوند.

داده کاوی بر روی یک سری اطلاعات کار ‌می‌کند که در نگاه اول احتمالاً کاربرد زیادی نخواهد داشت اما وقتی پژوهش‌های دقیق صورت می‌گیرد نتایج بدست آمده بسیار ارزشمند ‌می‌شود.

داده کاوی یک علم قوی است که ‌می‌تواند به همه بخش‌ها نفوذ داشته باشد، در دل اطلاعات برود و پاسخ به همه سوالاتی باشد که ایجاد ‌می‌شود. این علم در شرکت‌های بزرگ آنقدر مهم و شناخته شده است که بسیاری از برنامه‌ریزی‌ها بر اساس آن صورت ‌می‌گیرد. به عنوان مثال پیش از طراحی یک پروژه پرهزینه نیاز است که اطلاعات زیادی در مورد این پروژه مورد بررسی قرار گیرد تا بهترین تصمیم برای نحوه اجرای پروژه یا عدم اجرای آن گرفته شود.

تاریخچه

داده کاوي  فرآیند ي است که در آغاز دهۀ 90 مطرح شد و با نگرشی نو، به مسئلۀ استخراج اطلاعات از پایگاه داده ها میپردازد. از سال 1995 داده کاوي به صورت جدي وارد مباحث آمار شد و در سال 1996 ، اولین شمارة مجلۀ کشف دانش و معرفت از پایگاه داده ها  منتشر شد . محققانی نظیر براچمن و آناند ( 1996 ) کلیه مراحل واقع گرا یانه و رو به جلو کشف دانش از پایگاه داده ها را تشخیص دادند.

تا به امروز مفاهیم و کلمات زیادی برای علم داده‌کاوی گذاشته شده‌است که هرکدام منظور اینکار را تا حدودی یدک می‌کشیدند اما به اندازه این کلمه جامع و دربرگیرنده همه‌چیز نبودند. کارشناسان آمار ابتدا در سال ۱۹۶۰ که هنوز اسم مرسومی وجود نداشت، اصطلاح صید داده یا [Data Fishing](https://searchdatamanagement.techtarget.com/definition/data-dredging) و لای‌روبی داده یا [Data Dredging](https://en.wikipedia.org/wiki/Data_dredging) را برای این دانش و فرآیند در نظر گرفته بودند. تقریبا در سال  ۱۹۹۰ بود که کلمه‌ی “داده‌کاوی” در بین اساتید و کارکنان این علم جا باز کرد و خیلی زود به محبوبیت رسید. هنوز هم کلمات دیگری استفاده می‌شود که میتواند همین منظور را تا حدودی برساند مانند : یادگیری آماری، تحلیل داده و علم داده.

ضرورت و اهمیت داده کاوی

اهمیت علم داده کاوی زمانی بیشتر مطرح می شود که با حجم زیادی از داده ها مواجه شویم و این موضوع در همه ی منابع مربوط به این علم مورد تاکید قرار گرفته است.

به هر میزان حجم داده ها بیشتر باشد و رابطه ی موجود میان آنها پیچیده تر، دسترسی به اطلاعات و روابط پنهان میان داده ها دشوار تر می گردد.

در این بین نقش علم ” Data Mining ” به عنوان یکی از روش های تولید دانش از داده ی خام گویا تر می گردد. داده کاوی به صورت همزمان و تلفیقی از دانش هایی که شامل تکنولوژی پایگاه داده، یادگیری ماشین، هوش مصنوعی، شبکه های عصبی، آمار، الگو، سیستم های مبتنی بر دانش، بازیابی اطلاعات، حصول دانش، بازنمایی بصری داده و محاسبات سرعت بالا است، بهره می برد.

با توجه به توضیحات داده شده، داده‌کاوی بسیار مهم و کاربردی می‌باشد. پس در نتیجه باید به چشم فعالیت آینده‌دار به آن نگاه کرد. دیر یا زود هر کسب‌و‌کار و سازمانی متوجه میشود که باید داده‌کاوی را وارد سیستم خود کند و مقدار داده‌ای که استفاده چندانی از آن نمیشود، تبدیل به اطلاعات مفید شود. بعضی مواقع تنها برای اینکه سود بیشتری داشته باشیم، نیاز به داده‌کاوی در سازمان حس نمیشود. کلاهبرداری‌هایی که ممکن است در سیستم فروش اتفاق بیافتد، از طریق علم داده مشخص می‌شود.

کاربرد های داده کاوی و موارد استفاده از آن

از مهمترین کاربردهای داده کاوی، تحلیل و بررسی داده‌های صنایعی است که تمرکز بسیار زیادی روی مصرف‌کننده یا مشتریان دارند. تمامی صنایع دنیا می‌توانند از داده کاوی برای پیش‌بینی رفتار بازار و مصرف‌کننده و همچنین تاثیر سیاست‌گذاری‌های گوناگون بر بهبود عملکرد خود کمک بگیرند. صنایع مهندسی، پزشکی، محیط زیست، حمل و نقل و … همه و همه می‌توانند با استفاده از تکنیک‌های متنوع داده کاوی‌ همچون استفاده از درخت تصمیم‌گیری، شبکه‌ی عصبی، تحلیل [هوش مصنوعی](https://www.karlancer.com/blog/artificial-intelligence-threat/) و دسته‌بندی الگوها از مزایای این روش علمی بهره‌مند شوند.

**پزشکی**

یکی از مهم‌ترین کاربردهای داده کاوی، استفاده از آن در علم پزشکی است. حوزه‌ی پزشکی دارای شاخه‌های بسیار متنوعی است که هر یک از آن‌ها بر پایه داده‌های علمی و تجربی بنا شده‌اند و منبعی غنی از اطلاعات به شمار می‌آیند. تعیین نوع درمان یکی از متداول‌‌ترین کاربردهای داده ها در  صنعت پزشکی است. داده کاوی در این حوزه با کمک یک نمودار درختی انجام می‌شود که مناسب‌ترین انواع درمان را به بیمار پیشنهاد می‌دهد. کنترل عفونت‌های بیمارستانی، تشخیص عوارض جانبی داروها و همچنین تعیین متد جراحی جزو مواردی هستند که داده کاوی با کمک نمودار درختی و تعیین الگوریتم‌ها به پیش‌بینی الگوهای نهان آن می‌پردازد و به این ترتیب موجب پیشرفت صنایع مرتبط با طب و افزایش رضایت مراجعین می‌شود.

#### **جراحی‌ها**

کاربرد داده کاوی در پزشکی، در بخش جراحی نیز مشهود است. در این بخش متخصصین با استفاده الگوهای تجربی و داده‌های پیشین مواردی همچون سبک جراحی، میزان خون مورد نیاز برای تزریق، روش بیهوشی و مواردی از این دست را موردتحلیل و بررسی قرار می‌دهند و با این روش مناسب‌ترین و ارزان‌ترین راهکار را برای بیماران خود انتخاب می‌کنند.

### حفاظت از محیط زیست و منابع طبیعی و سیاره زمین

اهمیت استفاده از تکنولوژی برای حفظ محیط زیست و منابع طبیعی روز به روز نمایان تر می شود. محیط زیست به عنوان بزرگترین دارایی طبیعی نقش اساسی در کیفیت زندگی بشر دارد. تحلیل داده ها با استفاده از تکنولوژی های پیشرفته [ماشین لرنینگ](https://www.karlancer.com/blog/machine-learning/) می‌تواند با بررسی راهکارها و اطلاعات مربوط به منابع طبیعی و محیط زیست، به توسعه‌ی پایدار این منابع نیز کمک کند.

### تامین انواع انرژی برای خانه‌ها و ساختمان‌ها و صنایع

کاربرد داده کاوی در صنعت برق بخش‌های متنوعی را شامل می‌شود. این روزها با افزایش جمعیت و پیشرفت صنعت، مصرف انرژی نیز شکل فزاینده‌ای به خود گرفته است و قطعی‌های گاه و بی‌گاه برق یا سایر منابع انرژی نشان‌دهنده‌ی خارج شدن این مصرف از رویه‌ی نرمال است. تحلیل اطلاعات در این بخش می‌تواند با بررسی الگوی مصرف مشترکین مواردی همچون مصرف غیرمجاز را تشخیص دهد و با شناسایی مناطق بحران و پیش‌بینی نوسانات ولتاژ، خوشه‌بندی و تحلیل کارآمد را به موسسات توزیع نیرو ارائه دهد.

**بازاریابی و تبلیغات**

کاربرد داده کاوی در بازاریابی بسیار جذاب و جالب توجه است. این سیستم تحلیلی در زمینه‌ی بازاریابی به شما اجازه می‌دهد تا با تجزیه و تحلیل سبد خرید مشتریان خود و یافتن اقلام موردعلاقه‌ی آن‌ها، کالاهای مرتبط با خریدشان را به آن‌ها پیشنهاد بدهید و به این ترتیب فروش خود را افزایش دهید. داده کاوی در این شرایط از مقایسه‌‌ی نتایج گوناگون فروشگاه‌های مختلف و مشتریان گروه‌‌های مختلف، الگوی رفتار خرید مصرف‌کنندگان را شناسایی می‌کنند. تجزیه و تحلیل ‌های داده ها در مدیریت رابطه با مشتریان و طراحی [نرم افزار ارتباط با مشتریان](https://www.karlancer.com/blog/what-is-crm/) نیز کمک کننده است. [مدیریت ارتباط با مشتریان](https://www.karlancer.com/blog/what-is-crm/) نقش به‌سزایی در افزایش مشتریان رضایتمند و وفادار دارد و داده کاوی در این بخش کمک می‌کند تا با تحلیل علاقه‌مندی‌های مشتری خدمات و کالاهای مرتبط‌تری را ارائه دهید.

**حمل و نقل**

از دیگر کاربردها  استخراج قوانین و علم مرتبط با ترافیک و حمل و نقل است. صنعت حمل و نقل یک صنعت بسیار گسترده است که در سطوح بسیار بالاتر خود نیازمند استفاده از تکنیک‌های پیشرفته هوش مصنوعی و تحلیل اطلاعات  است. به این ترتیب مدیران صنایع حمل و نقل می‌توانند با داده کاویِ داده‌‌های بزرگ ترافیکی، به قوانین و علومی دست پیدا کنند که می‌تواند حجم عظیمی از رفت و آمدها در محیط حقیقی را کنترل کند. همین موضوع به کاهش تصادفات و حوادث مرتبط با حمل و نقل منجر می‌شود.
از سوی دیگر، بررسی مسیرهای عبور و مرور و تعیین برترین مسیرها با کمک داده های گذشته و فعلی، یکی از راهکارهای ارزنده در ترسیم نقشه‌های آفلاین و آنلاین است.

**کسب و کارهای آنلاین**

کاربرد داده کاوی در کسب و کارهای آنلاین نیز همچون کاربرد این متد در بازاریابی از الگویی یکسان پیروی می‌کند و به صاحبین کسب و کار آنلاین کمک می‌کند تا با تحلیل اطلاعات مربوط به کاربران و مشتریان خدمات و محصولات مناسب‌تری را در اختیار آن‌ها قرار دهند و به این وسیله فروش خود را تضمین کنند. اگر شما اقدام به [طراحی فروشگاه اینترنتی](https://www.karlancer.com/ecommerce-website) کرده اید و در صدد افزایش فروش با افزایش سبد خرید مشتریان کنونی هستید تحلیل اطلاعات خریداران و بازدید کنندگان محصولات می تواند به شما کمک شایانی کند.

**آموزش**

یکی از کاربردهای داده کاوی در زمینه‌ی آموزش است. آموزش به حدی برای این سیستم اهمیت دارد که یک شاخه‌ی مجزا از داده کاوی به نام داده کاوی آموزشی به آن اختصاص یافته است. این شاخه با متدهای توسعه‌ای سرو کار دارد که از داده‌‌های حاصل از محیط‌های آموزشی به وجود آمده‌اند. هدف اصلی داده کاوی آموزشی، پیش‌بینی رفتار یادگیری دانش آموزان در آینده، مطالعه‌‌ي تاثیر حمایت آموزشی و پیشرفت تکنولوژی‌های مرتبط با یادگیری است. معمولا موسسات آموزشی از داده کاوی آموزشی برای پیش‌بینی نتایج دانش‌آموزان و اخذ تصمیمات حساب شده استفاده می‌کنند. به این ترتیب آموزشگاه‌‌ها می‌توانند روی مباحث آموزشی و مدل آموزشی خود متمرکز شوند و با توجه به الگوهای یادگیری دانش‌آموزان خود، تکنیک‌های آموزشی‌شان را اتقا دهند.

**بانکداری**

بانکداری و سرمایه‌گذاری امروزی به شکل گسترده‌ای با خدمات آنلاین همراه شده‌اند و همین موضوع کمی کنترل اوضاع را سخت می‌کند؛ اما داده کاوی در این بخش هم به داد بشر رسیده است. کاربرد داده کاوی در بانکداری و سرمایه‌گذاری چندان پیچیده نیست. درواقع داده کاوی با یافتن الگوها، روابط علت و معلولی و همچنین همبستگی‌های پنهان اطلاعات مربوط به بازار و کسب‌وکار، مشکلات مالی بانکداری و موسسات مالی را به‌راحتی حل می‌کند. مدیران این بخش‌ها با در دست داشتن این اطلاعات و تحلیل‌ها می‌توانند بخش‌بندی و هدف‌گیری بازار را به شکل بهتری انجام دهند و در جذب و حفظ مشتریان خود موفق‌تر عمل کنند.

**شناسایی مجرمین**

دستگیری افراد متهم کار بسیار ساده‌ای است اما متاسفانه تشخیص مجرمین و پیدا کردن حقیقت اغلب دشوار است. داده کاوی در این زمینه نیز به کمک بشر می‌آید. مامورین می‌توانند با استفاده از تکنیک‌‌های داده کاوی به خصوص کاوش متن حقیقت را از میان دروغ بیرون بکشند. در این روش، داده کاوی از میان یک محتوای بدون ساختار الگویی ثابت و معنادار را بیرون می‌کشد. معمولا این الگوها با داده‌های جمع‌آوری شده از تحقیقات قبلی مقایسه می‌شود و یک مدل کلی برای سنجش دروغ ایجاد می‌کند.
یکی دیگر از کاربردهای داده کاوی استفاده از آن در جرم‌شناسی و یافتن جرائم و ارتباط آن با مجرمین است. داده‌های بسیار زیاد مرتبط با جرم و پیچیدگی روابط بین مجرم و جرم، باعث شده است تا داده کاوی یکی از مناسب‌ترین متدها برای یافتن الگوها در این حوزه باشد.

**کتاب‌داری و کتابخانه‌ها**

کتابخانه و کتاب‌داری یکی دیگر ازحوزه‌هایی هستند که با تغییر سبک زندگی بشر و روی کار آمدن تکنولوژی دست‌خوش تغییرات زیاد و مواجهه با گسترش داده‌ها شده‌اند. کاربردهای داده کاوی در کتابخانه‌ها بسیار گسترده است و فعالان این حوزه از داده کاوی برای طبقه‌بندی کامپیوتری کتاب‌ها، بررسی مرتبط‌ترین و جامع‌ترین لینک‌های منبع، حذف مطالب و کتاب‌های غیرمرتبط باعلائق مخاطبین، تولید خلاصه‌ی کتاب، خوشه‌بندی کتاب‌ها، کاهش ساعات کاری مسئولین کتابخانه و همچنین تقویت سیستم رفرنس‌دهی استفاده می‌کنند.

مزایا و معایب (کمبود ها) در این حوزه

### مزایا

داده کاوی دارای مزایایی می‌باشد که در زیر به چند مورد از آنها اشاره می‌شود:

**تصمیم گیری خودکار**

داده کاوی به سازمانها اجازه می‌دهد تا داده‌ها را به طور مداوم تجزیه و تحلیل کنند و تصمیمات مهم و عادی را بدون تأخیر در قضاوت انسان به طور خودکار انجام دهند. بانک‌ها میتوانند فوراً معاملات جعلی را تشخیص دهند، درخواست بازبینی کنند و حتی امنیت اطلاعات شخصی مشتریان در مقابل سرقت هویت آنها را تضمین کنند. این مدل‌ها در الگوریتم‌های عملیاتی یک شرکت مستقر هستند که می‌توانند داده‌ها را بطور مستقل جمع آوری و تجزیه و تحلیل کنند و به کار بگیرند تا تصمیم گیری را ساده تر و فرآیندهای روزانه یک سازمان را ارتقا دهند.

**پیش گویی صحیح و پیش بینی برنامه**

ریزی یک فرآیند مهم در هر سازمان است. داده کاوی برنامه ریزی را تسهیل می‌کند و بر اساس روند گذشته و شرایط فعلی پیش‌بینی‌های قابل قبولی را در اختیار مدیران قرار می‌دهد. داده کاوی مدل‌های پیش بینی تقاضا را پیاده‌سازی می‌کند، پیش‌بینی تقاضا برای محصولات فروشگاه‌ها و مسیریابی موجودی مناسب برای تأمین نیازهای بازار را انجام می‌دهد. بدین شکل داده‌کاوی به افزایش درآمد آن‌ها کمک می‌کند.

**کاهش هزینه**

داده کاوی امکان استفاده کارآمد و تخصیص منابع را فراهم می‌کند. سازمان‌ها می‌توانند با پیش‌بینی‌های دقیق تصمیم گیری‌های خودکار را برنامه ریزی و اتخاذ کنند که منجر به کاهش هزینه می‌شود.

**بهینه سازی محصولات و خدمات**

شناخت محصولات پرفروش، محصولات سودآور، محصولات زیان ده و… از جمله فواید داده کاوی است که شما را به عنوان مدیر کسب و کار برای افزایش کیفیت محصولات ترغیب می‌کند.

**شناخت مشتریان وفادار**

با وجود داده کاوی می‌توانید بفهمید مشتریان قدیمی شما چه کسانی هستند، چه کالاهایی را دوست دارند، چه برنامه‌ای برای خرید دارند، کدام کالاها باعث وفاداری آن‌ها شده است، و…

**شناسایی و بررسی رفتار مشتری**

سعی کنید رفتار مشتریان خود را بشناسید و آن را با ویژگی‌های او مطابقت دهید. اگر بتوانید این کار را انجام دهید، در زمینه بخش بندی بازار موفق‌تر عمل خواهید کرد.

**دیدگاه مشتری**

این داده‌ها می‌توانند بینش شرکتها را تغییر دهند به طوریکه شرکته‌ا مدل‌های داده کاوی را بر اساس داده‌های مشتریان گسترش می‌دهند و به شناسایی خصوصیات و تفاوت‌های کلیدی در بین مشتریان خود می‌پردازند. این بینش باعث نگهداری مشتری و سوددهی شرکت می‌شود.

**بررسی چرخه عمر مشتری**

با استفاده از داده کاوی می‌توانید چرخه عمر مشتری و همچنین میزان سود به دست آمده از مشتری در هر مرحله را بررسی کنید.

داده کاوی در زمینه‌های علمی‌و سلامت و سیاست و همینطور اقتصاد نیز دارای مزایای زیادی است. در دنیای سخت و بی‌رحم کرونایی، داده کاوی توانست الگوهای جدید دارویی را کشف نماید. داده کاوی در دنیای بورس نیز وارد شده است و توانسته رشد خوبی ایجاد نماید.

### معایب

امنیت مسئله بزرگی است. مشاغل دارای اطلاعاتی در مورد کارمندان و مشتریان خود از جمله شماره تأمین اجتماعی، روز تولد، حقوق و دستمزد و غیره هستند. با این وجود چگونگی مراقبت صحیح از این اطلاعات هنوز جای سوال دارد. موارد زیادی وجود داشته است که هکرها با داشتن اطلاعات شخصی و مالی بسیار زیاد، کارت اعتباری سرقت شده و سرقت هویت به یک معضل بزرگ دسترسی پیدا کرده و داده‌های کلیدی مشتریان شرکت بزرگ مانند شرکت اعتبار فورد موتور ، سونی را دزدیده‌اند.

اطلاعات جمع آوری شده از طریق داده کاوی در نظر گرفته شده برای اهداف اخلاقی می‌تواند مورد سوءاستفاده قرار گیرد که از معایب داده کاوی به شمار می‌رود. ممکن است افراد غیراخلاقی یا مشاغل از این اطلاعات برای بهره‌مندی از افراد آسیب پذیر یا تبعیض علیه گروهی استفاده کنند..

مراحل و استراتژی های موجود در داده کاوی

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» شوند. مرحله آماده‌سازی داده‌ها قبل از پردازش را، پیش‌پردازش (Preprocessing) می‌گویند. 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، در ادامه بیان شده‌اند.

مراحل پیش پردازش داده عموما به دو دسته تقسیم میشوند که عبارتند از:

* انتخاب اشیا داده و ویژگی ها (attribute) برای تحلیل
* ایجاد کردن یا تغییر دادن ویژگی ها

مهمترین فعالیت های که در بخش پیش پردازش داده‌ها انجام می شود عبارت است از:

1. پاکسازی داده (data cleaning)
2. یکپارچه سازی داده (data integration )
3. کاهش داده (data reduction)
4. تبدیل داده (data transformation)
* **پاکسازی داده (data cleaning)**

مهمترین فعالیت های این بخش عبارت است تخمین مقادیر ناموجود در پایگاه داده ها، از بین بردن اختلال (noise) در داده ها، حذف کردن داده های پرت و نامربوط، از بین بردن ناسازگاری در داده ها. نکته مهم این است که هر چه این گام از داده کاوی بهتر انجام شود، خروجی الگوریتم ها و تکنیک های داده کاوی کیفیت بالاتری خواهد داشت.

* **یکپارچه سازی داده (data integration)**

در بسیاری از موارد ممکن است داده ها در فایل ها و منابع مختلف نگهداری شوند و در این صورت نیاز است تا داده ها پیش از اجرای تکنیک های داده کاوی با یکدیگر یکپارچه شوند.  یکپارچه سازی هم فعالیتی سنگین است و هم چالش های فراوانی را به همراه دارد که در مفصل در مورد آن صحبت خواهیم کرد.

* **کاهش داده (data reduction)**

یکی از مهمترین نکات در داده کاوی این است که ممکن است همیشه، همه داده ها مورد نیاز نباشند و تنها بخشی از داده ها که مورد نیاز است باید مورد پردازش قرار بگیرد. کاهش داده (data reduction) به این مباحث می پردازد.

* **تبدیل داده (data transformation)**

فعالیت های مانند نرمال سازی داده ها و گسسته سازی داده ها در این حوزه جای میگیرند

**مراحل داده کاوی**

**معمولاً داده‌کاوی در شش گام زیر صورت می گیرد:**

**درک کسب و کار**

**بررسی و درک داده ها**

**آماده سازی داده‌ها**

 **مدل‌سازی**

**تست و ارزیابی مدل**

**توسعه مدل نهایی و استقرار**



**شکل1**

**گام اول: درک کسب‌وکار**

یکی از مراحل مهم یک پروژه داده‌کاوی فهم نیاز کسب‌وکار است. این کار با [مطالعه و فهم دقیق نیازهای مدیریتی](https://analica.ir/business-analysis/) آغاز می‌شود. اهداف کسب‌وکار که انگیزه اصلی اجرای پروژه است باید به‌خوبی مشخص شوند. اهدافی مانند این‌که “ویژگی‌های مشترک مشتریانی که اخیراً از دست دادیم و از خدمات و محصولات شرکت‌های رقیب استفاده می‌کنند، چیست؟” یا “هر یک از مشتریان شرکت دارای چه ارزشی برای ما هستند؟” .

فهم کسب‌وکار و هدف اصلی اجرای پروژه مشخص می‌کند چه داده‌هایی باید جمع‌آوری شوند، چگونه داده‌ها تحلیل شوند و چطور نتایج ارائه شوند. همچنین کمک می‌کند تا بودجه موردنیاز برای اجرا و زمان‌بندی پروژه تعیین گردد.

**گام دوم: بررسی و درک داده**

با توجه به نیاز کسب‌وکار، مجموعه‌ای از داده‌ها که می‌توانیم از آن‌ها استفاده کنیم تا هدف آن پروژه محقق گردد، شناسایی می‌شوند.

**گام سوم: آماده‌سازی داده ها**

هدف از این گام، آماده کردن داده‌ها برای فاز تحلیل با روش‌های داده‌کاوی است.



شکل2

شکل-۲ نشان می‌دهد که در یک پروژه داده‌کاوی چه مراحلی باید طی گردد تا داده‌های دنیای کسب‌وکار برای تحلیل نهایی آماده شوند.

**گام چهارم: مدل‌سازی**

در این گام، تحلیل‌گر ممکن است روش‌های مختلف داده‌کاوی را بر روی داده‌های آماده‌شده امتحان کند تا بتواند به هدف اصلی پروژه برسد. ساخت مدل یک فرآیند خطی نیست و رفت‌وبرگشت‌های زیادی وجود دارد. یک مدل بهینه در داده‌کاوی وجود ندارد و بسته به مسئله‌ای که تحلیل‌گر با آن مواجه است، روش‌های مختلف باید آزمایش شوند و خروجی آن‌ها باهم مقایسه گردند. در این مرحله احتمالاً لازم است به گام قبلی بازگشت و برای برخی از الگوریتم‌ها داده‌ها را به شکل دیگری آماده کرد.

**گام پنجم: تست و ارزیابی مدل**

در گام پنجم مدلی که توسعه یافته است بر اساس دقت و قابلیت عمومی‌سازی آن آزمایش می‌شود. در این مرحله باید ارزیابی شود که مدل تا چه حد می‌تواند به اهداف کسب‌وکار کمک کند. اگر زمان‌بندی و بودجه پروژه اجازه دهد بهتر است مدل در دنیای واقعی آزمایش شود. نتایج آزمایش کمک می‌کند تا مدل ارزیابی شود و شاید اطلاعات جدیدتری به دست آید که به کامل‌تر شدن مدل کمک کند.

**گام ششم: توسعه مدل نهایی و استقرار**

بسته به نوع پروژه، فاز استقرار می‌تواند متفاوت باشد. در برخی موارد ارائه گزارش از روند کار و خروجی تحلیل، پایان یک پروژه داده‌کاوی است. در سمت دیگر استقرار یک سیستم قابل تکرار که سازمان از آن بتواند برای مدت‌ها استفاده کند قرار دارد. در استقرار چنین سیستمی باید تحلیل‌گر نیز مشارکت داده شود تا فهم خود را به اجراکننده سیستم انتقال دهد.

مرحله استقرار می‌تواند شامل فعالیت‌های نگه‌داری نیز شود. در طول زمان محیط کسب‌وکار و نیازهای آن تغییر می‌کند و ممکن است مدل به‌دست‌آمده کارایی خود را از دست بدهد. طراحی یک استراتژی نگه‌داری مناسب می‌تواند کمک کند تا کسب‌وکار برای مدت طولانی به‌اشتباه از مدل داده‌کاوی استفاده نکند.

**استراتژی های موجود در داده کاوی**

استراتژی های مختلف داده كاوی را می توان بر اساس نوع عملیاتی كه انجام می دهند به دو دسته « پیش بینی كننده » و « تشریح كننده » تقسیم كرد . استراتژی های پیش بینی كننده با ساخت مدلی برای پایگاه داده ، وظیفه پیش بینی موارد ناشناخته را بر عهده دارند . در حالی كه استراتژی های تشریح كننده ، الگوهایی قابل فهم از داده ها را برای انسان كشف می كنند .

ابزار ها و روش های موجود در این حوزه

#### **Orange**

یک نرم افزار متن باز و رایگان است که در پایتون نوشته شده است. یکی از بهترین برنامه ها برای تجزیه و تحلیل و داده کاوی پایه است. Orange با یک رابط کاربری منحصر به فرد و کاربر پسند، رویکردی کاربرگرا را برای داده کاوی در نظر گرفته است.

Rapid Miner

این نرم افزار که به زبان جاوا نوشته شده یکی از برترین نرم افزارها برای انجام تجزیه و تحلیل و پیش بینی است که محیط های یکپارچه ایی برای یادگیری عمیق، استخراج متن و یادگیری ماشین ارائه می‌دهد. این پلتفرم می‌تواند از سرورهای ثابت یا مبتنی بر رایانش ابری استفاده کند و در مجموعه‌های متنوعی از سازمان‌ها پیاده سازی شده است.

Mahout

یک نرم افزار متن باز است که برپایه فرآیند یادگیری بدون نظارت متمرکز شده است. این نرم افزار در ایجاد الگوریتم‌های یادگیری ماشین برای خوشه بندی، طبقه بندی و فیلترهای مشترک  مزایای زیادی دارد. این برنامه به ریاضیدانان، آمارشناسان و دانشمندان داده اجازه می‌دهد تا الگوریتم های خاص خود را ایجاد، آزمایش و پیاده سازی کنند. در حالی که Mahout مانند یک توصیه کننده عمل می‌کند، که سازمان‌ها می توانند با حداقل تلاش از قابلیت های کامل آن استفاده کنند.

Micro Strategy

یک نرم افزار تجزیه و تحلیل اطلاعات کسب و کار و داده است که تمام مدل‌های داده کاوی را دارد. این پلت فرم با داشتن طیف گسترده‌ای از دروازه‌ها و درایورهای بومی، می‌تواند به هر منبع اطلاعات سازمانی متصل شده و داده های آن را تجزیه و تحلیل کند. این نرم افزار می‌تواند عملکرد همه مدل‌های داده کاوی را در زمان واقعی ردیابی و تجزیه و تحلیل کند و نتایج را بصورت دیدگاه‌های قابل بررسی برای تصمیم گیرندگان، به وضوح نشان دهد.

#### IBM SPSS Modeler

اگر در مقیاس گسترده‌ای از پروژه‌هایی مانند آنالیز متنی کار می‌کنید حتما میز کار IBM SPSS و رابط تصویری آن را کشف خواهید کرد. این ابزار شما را قادر می‌سازد طیف گسترده‌ای از الگوریتم‌های داده کاوی را بدون داشتن اطلاعاتی در مورد برنامه نویسی تولید کنید.

#### Tableau

یک برنامه نویسی داده تجسمی است که با ایجاد گرافیک شگفت انگیز برای ایجاد تصویری تعاملی طراحی شده است. این ابزار داده کاوی مناسب صنایعی است که در زمینه هوش تجاری (Business Intelligence) فعالیت می‌کنند.

#### Weka

شامل مجموعه‌ای از ابزارها و الگوریتم‌های تجسم برای تجزیه و تحلیل داده‌ها و مدل‌سازی پیش بینی، همراه با رابط‌های گرافیکی کاربر برای دسترسی آسان به این توابع است.

نرم افزار Weka یک نرم افزار اوپن سورس جهت پیاده سازی الگوریتم‌های یادگیری ماشین و داده کاوی است. از این نرم افزار برای حل مسائل داده کاوی در دنیای واقعی استفاده می‌شود. نرم افزار داده کاوی وکا به زبان جاوا نوشته شده است و تقریبا روی همه سیستم عامل‌ها و پلتفرم‌ها قابل اجراست.

وکا یک نرم افزار جامع است که به شما امکان می‌دهد، داده‌های بزرگ (Big data) را پردازش کنید، الگوریتم‌های مختلف داده کاوی را روی آن‌ها اعمال کنید و خروجی‌های مختلف را با یکدیگر مقایسه کنید. این نرم افزار کار با داده‌های بزرگ را آسان کرده و امکان یادگیری ماشینی با استفاده از الگوریتم‌های مختلف را فراهم می‌کند.

نرم افزار وکا در ابتدا با اهداف تحقیقاتی در دانشگاه دانشگاه وایکاتو (University of Waikato) کشور نیوزیلند، ایجاد و توسعه داده شد، اما بعدها در اختیار همگان قرار گرفت و امروزه محققان، کارشناسان و متخصصان علم داده در سراسر جهان و در حوزه‌های مختلف از آن استفاده می‌کنند.

نام Weka از عبارت «Waikato Environment for knowledge Analysis» استخراج شده است. همچنین Weka نوعی پرنده‌ در کشور نیوزیلند است که توانایی پرواز کردن ندارد اما در قسمت‌های مختلف طبیعت به جستجو می‌پردازد.

در شکل زیر ویژگی های **Weka** آمده است :



شکل 3

## **امکانات نرم افزار وکا**

نرم افزار Weka از کل فرآیند داده کاوی و یادگیری ماشین از آماده سازی داده‌های ورودی، اعمال الگوریتم و ارزیابی داده‌ها تا مصور سازی و نمایش گرافیکی نتایج پشتیبانی می‌کند. برخی از امکانات این نرم افزار شامل موارد زیر است.

* پیش پردازش (Preprocessing)
* دسته بندی (Classification)
* رگرسیون (Regression)
* خوشه بندی (Clustering)
* قوانین انجمنی (Association rules)
* انتخاب ویژگی (Attribute selection)
* آزمایش (Experiments)
* گردش کار (Workflow)
* مصورسازی (Visualization)

### کاربردهای نرم افزار Weka

تقریبا در کلیه حوزه‌های یادگیری ماشین و داده کاوی می‌توان از نرم افزار Weka استفاده کرد:

* **آموزشی:** دانشگاه‌ها و مراکز آموزش علم داده از این نرم افزار برای اهداف آموزشی استفاده می‌کنند.
* **تحقیقاتی:** بسیاری از محققان حوزه علم داده از قابلیت‌های این نرم افزار جهت بررسی پژوهش‌های خود بهره می‌برند.
* **کاربردهای تجاری و صنعتی:** برای حل بسیاری از مسائل واقعی مرتبط با داده‌ها از این نرم افزار استفاده می‌شود.

**بخش های اصلی Weka**

 آسان­ترين راه استفاده از Weka ، از طريق واسطي گرافيكي است كه Explorer خوانده مي‏شود. اين واسط گرافيكي، به وسيله انتخاب منوها و پر كردن فرم­‏هاي مربوطه، دسترسي به همه امكانات را فراهم كرده است. براي مثال، مي‏­توان به سرعت يك مجموعه داده را از يك فايل ARFF خواند و درخت تصميم‏گيري آن را توليد نمود. امادرخت­‏هاي تصميم‏گيري يادگيرنده صرفاً ابتداي كار هستند. الگوريتم­‏هاي بسيار ديگري براي جستجو وجود دارند. واسط Explorer كمك مي‏­كند تا الگوريتم­‏هاي ديگر نيز آزمايش شوند.



شکل 4- Weka در وضعيت انتخاب واسط

اين شامل چهار واسط کاربری متفاوت می‌باشد:

**Explorer -1**

اين واسط گرافيکي، محيطي برا مکاشفه در داده ها با استفاده از WEKA مي باشد، که به وسيله انتخاب منوها و پر کردن فرم های مربوطه، دسترسي به همه امکانات را فراهم کرده است. واسط Explorer کمک مي کند تا الگوريتم ها متعدد نيز آزمايش شوند.

در اين واسط، شش پانل مختلف وجود دارد. کارکرد تمام گزينه ها به شرح ذيل مي باشد:



شکل 5 - واسط گرافيکي Explorer

: Preprocess انتخاب مجموعه داده و اصلاح آن از راه ها گوناگون

Classify: آموزش برنامه ها يادگير که رده بند يادگير رگرسيون انجام مي دهند و ارزيابي آن ها

 Cluster: يادگير خوشه ها براƽ مجموعه ها داده.

Associate: يادگير قواعد انجمني برا داده ها و ارزيابي آن ها.

: Select Attributes انتخاب مرتبط ترين جنبه ها در مجموعه ها داده.

 Visualize :مشاهده نمودارهاƽ مختلف دو بعدƽ داده ها و تعامل با آن ها.

 **Experimenter-2**

 محيطي برا انجام آزمون و انجام آزمايش ها آمار ميان روش های مختلف يادگير مي باشد. اين واسط کمک مي کند تا به هنگام استفاده از تکنيک ها رده بند و رگريون، چه روش ها و پارامترهايي برا مساله داده شده، بهتر عمل مي کنند. با اين وجود، Experimenter با ساده کردن اجرا رده بند کننده ها و فيلترها با پارامترها گوناگون رو تعداد از مجموعه ها داده، جمع آورکارآيي و انجام آزمايش ها معنا، پردازش را خودکار مي کند. معمولا کاربرها پيشرفته، از Experimenter برا توزيع بار محاسباتي بين چندين ماشين، استفاده مي کنند.

**Knowledge Flow-3**

محيطي که تقريبا تمامي کارايي Explorer را پشتيباني مي کند، اما در عوض ١ از يک رابط «کشیدن و رھا کردن » استفاده می کنند که با اتصال الگوریتم ھای یادگیری و منابع داده ھا، ترکیب وچینش دلخواه ساختھ می شود.

**Simple CLI-4**

رابط ساده خط فرمان که به شما اجازه دسترسي مستقيم به دستورات WEKA را مي دهد. ما Explorer به عنوان واسط کاربر انتخاب مي نمائيم. در محيط Explorer WEKA ،در پايين هر پانل، جعبه Status و دکمه Log قرار دارد. جعبه Status ،پيغام ها مبني بر انجام عمليات را نشان مي دهد و دکمه Log ،گزارش عملکرد متني کارهايي که WEKA تاکنون در اين بخش انجام داده را به همراه برچسب زماني ارائه مي کند. توجه داشته باشيد زماني که WEKA در حال عمليات است، پرنده کوچکي در پايين سمت راست پنجره بالا و پايين مي پرد. اگر پرنده بايستد و حرکت نکند، او مريض است؛ يعني اشتباهي رخ داده است و بايد Explorer دوباره اجرا شود.

**برخی از توابع وکا به منظور آماده سازی داده‌ها به شرح زیر می باشد:**

* تبدیل متغیرهای گسسته چند مقداری به دو مقداری و تبدیل متغیرهای پیوسته به گسسته
* روش‌های نمونه گیری با جایگذاری و بدون جایگذاری و روش‌های پیشرفته تر مانند SMOTE
* گسسته سازی بدون نظارت و با نظارت
* نرمالسازی و استانداردسازی
* روش LOF (Local Outlier Factor) برای پیدا کردن نقاط دورافتاده
* روش‌های مختلف برای ادغام مقادیر مختلف متغیرهای گسسته
* جایگذاری مقادیر از دست رفته
* کاهش بعد داده ‌ها با استفاده از تحلیل اجزای اصلی (PCA) و موجک (Wavelet)

**روش‌های خوشه‌بندی در این نرم افزار نیز به شرح زیر هستند:**

* روش‌های بر پایه مرکز هندسی
* روش‌های بر پایه چگالی: DBSCAN و OPTICS
* روش‌های سلسله مراتبی
* روش‌های برپایه توزیع احتمالی مانند EM

**روش‌های کشف قواعد انجمنی نیز به شرح زیر هستند**:

* روش Apriori
* کشف قواعد انجمنی متوالی
* روش‌های درختی مانند FP growth

در نرم افزار وکا همچنین قابلیت کار با مجموعه داده‌های بزرگ و به اصطلاح Big Data وجود دارد. در هنگام مواجه با این مسائل بهترین پیشنهاد استفاده از پکیج command-line می‌باشد. همچنین امکان نوشتن کد مستقیما در جاوا یا زبان‌های بر پایه جاوا مانند Groovy یا Jython نیز وجود دارد.

محاسن این نرم افزار عبارتند از:

* در بردارنده محدوده وسیعی آماده سازی داده‌ها و روش‌های انتخاب ویژگی‌ها به صورت یکپارچه.
* محدوده وسیعی از روش های رده بندی
* وجود تعداد زیادی شاخص به منظور ارزیابی روش های رده بندی
* محیط های مختلف کاری به منظور سهولت در اجرای مدلسازی های مختلف

برخی از معایب آن نیز عبارند از:

* زمان‌بر بودن اجرای مدل‌ها به علت عدم بهینه بودن برخی از آنها (در برخی موارد زمان اجرای یک الگوریتم در این نرم افزار در مقایسه با نرم افزاری مانند IBM modeler حتی به 5 برابر نیز می رسد(
* محدودیت در منابع آموزشی
* محدودیت در روش های خوشه بندی و قواعد انجمنی
* محدودیت شدید در مصورسازی داده ها

**تحلیل نهایی:**

نرم افزار وکا به منظور اجرای مدل های رده بندی علی رقم سرعت پایین آن، یکی از حرفه ای ترنی نرم افزارهای موجود است. تعداد بیشماری از مقالات منتشر شده تاکنون از این نرم افزار به منظور مدلسازی های خود استفاده نموده اند. همچنین این نرم افزار با داشتن دامنه وسیعی از روش های انتخاب ویژگی بسیار مناسب برای تحقیقاتی است که می خواهند از این روش ها استفاده کنند. یک ضعف اصلی این نرم افزار عدم وجود تحلیل کافی در خصوص روش های خوشه بندی است و همچنین عدم وجود شاخص های ارزیابی برای این روش ها است. در نهایت وکا با وجود محیط های مختلف در کوتاهترین زمان قابل یادگیری است و این یکی از مزیت های اصلی آن نیز می باشد.

منابع و مراجع

* شهرابی، جمال .کتاب داده کاوي. موسسه پژوهشی داده پردازان گیتا و جهاد دانشگاهی واحد صنعتی امیر کبیر. چاپ اول. سال 1386.
* سعیدي، احمد . مقاله داده کاوي، مفهوم و کاربرد آن در آموزش عالی . فصلنامه نامه آموزش عالی . شماره 18 .اسفند 1384
* کیخا، مصطفی. عباسی، علی. مقدمهاي بر دادهکاوي
* ویدیو آموزشی داده کاوی از صفر تا صد مجموعه فرادرس
* <https://dayche.com/data-mining-definition/>
* <https://www.hamyarit.com/basic/data-mining/>
* <https://karokasb.org/data-mining/>
* <https://www.irandnn.ir/mag/what-is-data-mining/>
* Seifert,Jeffery.W.Data Mining : An Overview, December2004