مرکز تحقیقات فناوری اطلاعات در امور سلامت با همکاری معاونت بهداشتی دانشگاه علوم پزشکی اصفهان

داده کاوی در نظام سلامت

دکتر محمد ستاری

دکترای مهندسی کامپیوتر - نرم افزار

عضو هیات علمی گروه مدیریت و فناوری اطلاعات سلامت دانشگاه علوم پزشکی اصفهان

عضو مرکز تحقیقات فناوری اطلاعات در امور سلامت

فهرست مطالب

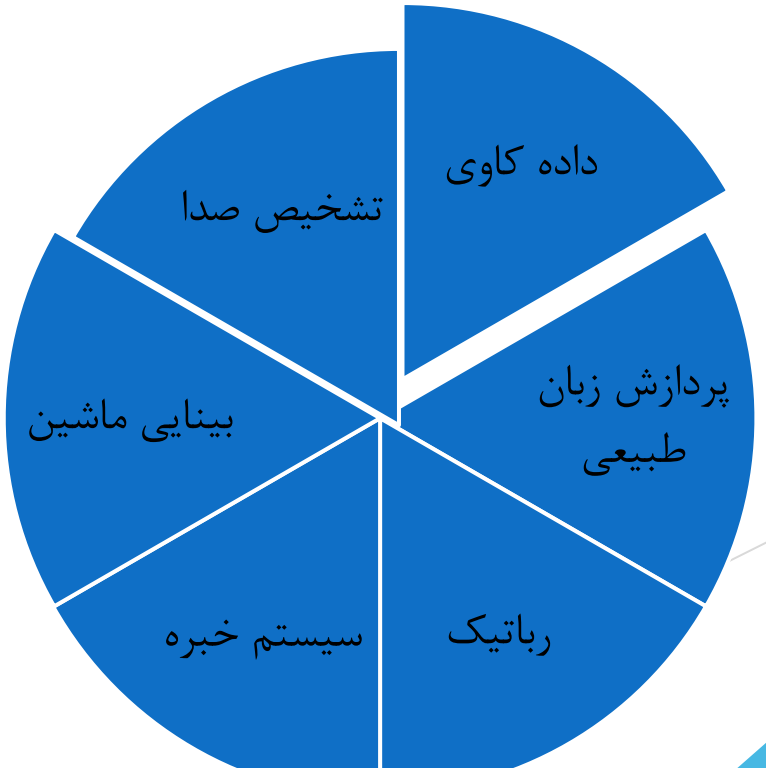
تاریخچه هوش مصنوعی

- شروع و سرچشمه هوش مصنوعی به سال های جنگ جهانی دوم بر می گردد. زمانی که نیروهای آلمانی برای رمز نگاری و ارسال ایمن پیام ها از ماشین enigma استفاده می کردند و دانشمند انگلیسی، آلن تورینگ در تلاش برای شکست این کدها برآمد. تورینگ به همراه تیمش ماشین bombe را ساختند که enigma را رمز گشایی می کرد. هر دو ماشین enigma و bombe پایه های یادگیری ماشینی machine learning هستند که یکی از شاخه های هوش مصنوعی یا همان Artificial intelligence می باشد.

- تورینگ ماشینی را هوشمند می دانست که بدون اینکه به انسان حس صحبت با ماشین را بدهد، با او ارتباط برقرار کند و این مسئله پایه علم هوش مصنوعی است یعنی ساخت ماشینی که همانند انسان فکر، تصمیم گیری و عمل کند.

هوش مصنوعی

هوش مصنوعی یک فیلد مطالعاتی است که سعی دارد قابلیت‌های استدلال، برنامه‌ریزی، حل مسئله و یادگیری را در کامپیوتر ایجاد نماید.



مقدمه (داده کاوی)

از هنگامی که رایانه در تحلیل و ذخیره سازی داده ها بکار رفت (۱۹۵۰) پس از حدود ۲۱ سال حجم داده ها در پایگاه داده ها دو برابر شد. ولی پس از گذشت دو دهه و همزمان با پیشرفت فن آوری اطلاعات، هر دو سال یکبار حجم داده ها دو برابر شده و همچنین تعداد پایگاه داده ها با سرعت بیشتری در حال رشد است

حال با ورود سیستم های یکپارچه اطلاعاتی، لحظه به لحظه به حجم داده ها در پایگاه داده های مربوط اضافه شده و باعث به وجود آمدن انبارهای عظیمی از داده ها شده است.

سازمان ها از لحاظ داده غنی از لحاظ دانش ضعیف

تعریف داده کاوی

- فلسفه ی داده کاوی این است که آینده بسیار به گذشته شبیه است.
- داده کاوی به شما کمک می کند تا رفتار کسب و کار خود در گذشته را دقیق بشناسید و بر اساس آن آینده را با تقریب بالایی پیش بینی کنید.

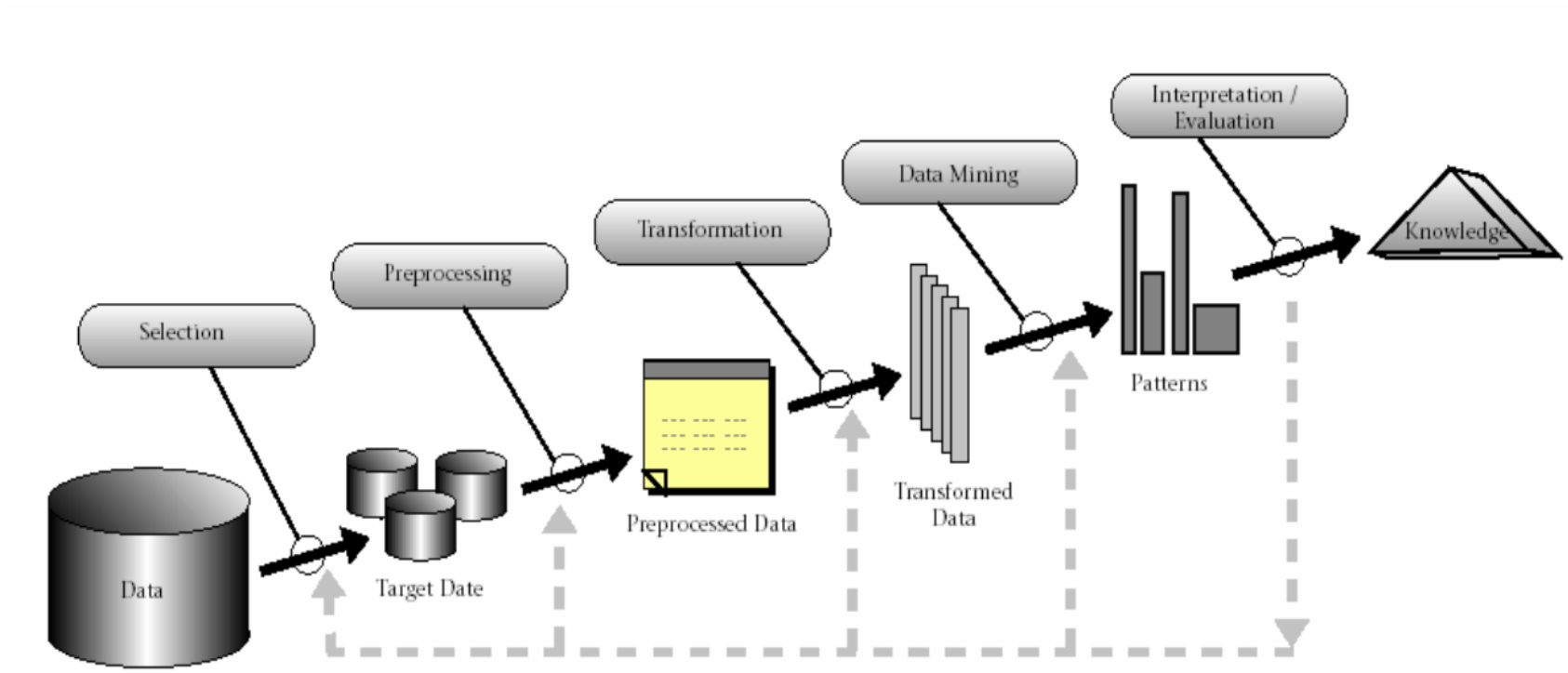
مطالعات داده کاوی

- مقطعی
- گذشته نگر

تعریف 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.

مراحل تبدیل داده به دانش



مشکلات قابل حل توسط داده کاوی

- مسئله ای پیچیده و ناساخت یافته و یا نیمه ساخت یافته
- داده‌های مرتبط وجود داشته باشند و به آنها دسترسی داشت
- داده ها در یکجا مجتمع شده و انباره داده ها ایجاد شود.
- توانایی کامپیوترها امکان استفاده از نرم افزارهای مرتبط با داده کاوی را به ما بدهند
- مدیران نیاز به استفاده از دانش استخراج شده از داده ها را حس کرده باشند

چرا داده کاوی

۹۰ درصد اطلاعات دنیای دیجیتال را داده‌های بدون ساختار (unstructured data) تشکیل می‌دهد.

منجر به **تصمیمات واقع بینانه** می شود.

سبب **تکرار تصمیمات سودآور** رخ داده در گذشته می گردد.

اهمیت داده کاوی در حوزه سلامت

امروزه صنعت بهداشت و درمان مقادیر زیادی داده پیچیده در مورد منابع بیمارستانی ، تشخیص بیماری ها ، پرونده های الکترونیکی بیماران ، تجهیزات پزشکی و غیره تولید می کند.

مقادیر زیادی از داده ها یک منبع کلیدی برای پردازش و تجزیه و تحلیل برای استخراج دانش است که پشتیبانی از صرفه جویی در هزینه و تصمیم گیری را امکان پذیر می کند.

کاربرد داده کاوی در حوزه سلامت

۱) تشخیص، درمان بیماری

۲) مدیریت منابع پزشکی

۳) جلوگیری از سو استفاده و تقلب

۴) کرونا

تشخیص، درمان بیماری

داده های سوابق درمانی خود را استخراج کرده است تا داروهای مناسب را کشف کنند

از طریق مطالعه ی گذشته نگر، می تواند الگوهای تشخیصی پرتکرار با میزان اطمینان بالا را استخراج نماید و این الگوها را با مشاوره با یک متخصص ارزیابی نماید و از این الگوها در جهت پیشگیری از بیماری ها استفاده نماید

مدیریت منابع پزشکی

- کاهش طول مدت اقامت بیمار
- جلوگیری از عوارض بالینی
- انتخاب بهترین روش ها
- ارائه اطلاعات مناسب به پزشکان
- کاهش هزینه ها

جلوگیری از سو استفاده و تقلب

اداره ی کلاهبرداری پزشکی یوتا با تحلیل انبوه داده های تولید شده توسط میلیون ها نسخه ، دوره عملیاتی و درمانی را برای شناسایی الگوهای غیرمعمول و کشف تقلب استخراج کرده است

شناسایی افرادی که سواستفاده انجام داده اند

کاربرد داده کاوی در کرونا

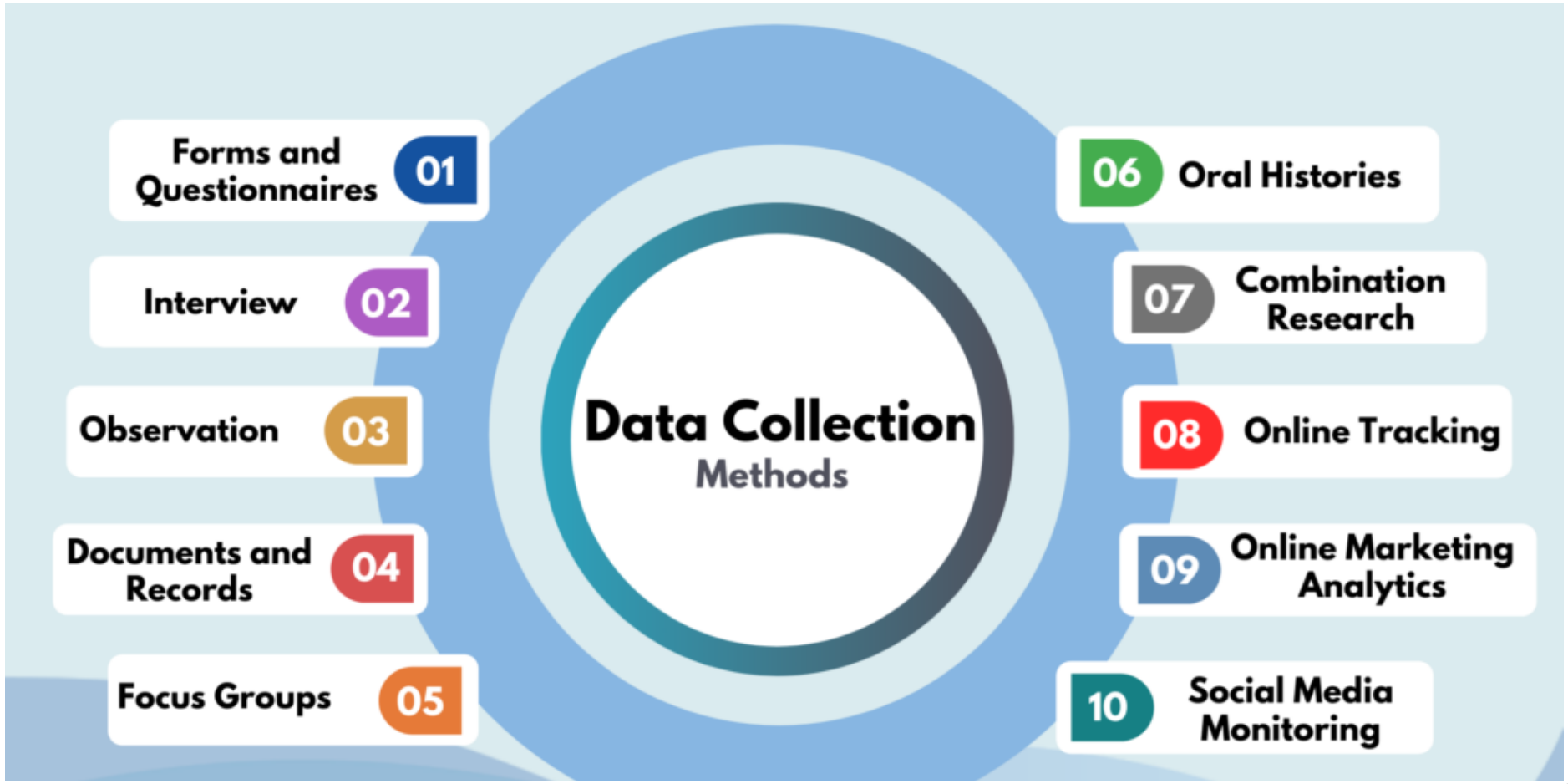
- 1)**پیش بینی آغاز شیوع:** از تکنیک های داده کاوی و یادگیری ماشین برای پیش بینی یا تخمین حالت های آینده یک سیستم یا محیط استفاده می‌شود. یکی دیگر از حوزه هایی که مدل سازی پیش بینی مبتنی بر هوش مصنوعی در آن پررنگ شده است، حوزه شناسایی بیماری های واگیرداری مانند ویروس کرونا است. به عنوان مثال شرکت بلودات واقع در کانادا به خوبی شیوع کووید ۱۹ را در دهم دی ماه ۱۳۹۸ در کشور چین پیش بینی و اعلام کرد.
- ۲) **پیش بینی و مدل سازی نحوه گسترش:** پژوهشگران دانشگاه ساوت همپتون نحوه شیوع این بیماری را در طی جشن های ۴۰ روزه ی سال نو که بیشترین مسافرت ها در آن رخ می‌دهد، بر اساس داده های انجمن بین المللی حمل و نقل هوایی و داده های مکان یابی با کمک الگوریتم های داده کاوی پیش بینی کردند.

کاربرد داده کاوی در کرونا

- **۳) تشخیص بیماری:** یکی از کاربردهای داده کاوی در تشخیص بیماری کرونا استفاده از سی تی اسکن ریه افراد است. این الگوریتم ها از دقت بالایی برخوردار است و با استفاده از تصاویر سی تی اسکن هزاران مورد تأیید شده مبتلا به کرونا ارائه شده است. سیستم به طور رسمی توسط شرکت علی بابا (غول تکنولوژیک چین) ارائه شد و در صدها بیمارستان به کار گرفته شد. یکی دیگر از موارد استفاده داده کاوی برای تشخیص ویروس کرونا، سامانه های اندازه گیری دمای بدن برای تشخیص تب است. دو نمونه از این سامانه ها توسط شرکت های بایدو و مگوی توسعه داده شده‌اند. هر دوی این سامانه ها از ترکیب هوش مصنوعی و دوربین های حرارتی برای شناسایی افراد مشکوک به کرونا در مترو و در معابر پرتردد استفاده می کنند. سامانه شرکت مگوی می‌تواند با خطای ۰,۳ درجه سلسیوس و در شعاع ۵ متری دمای افراد را حتی اگر ماسک و کلاه داشته باشند، اندازه گیری کند. این سامانه می تواند در هر ثانیه دمای بدن ۱۵ نفر را بسنجد و با استفاده از یک اپراتور انسانی افراد مشکوک را شناسایی کرده و اخطار بدهد.

متدولوژی crisp

روش های جمع آوری اطلاعات



مشکلات جمع آوری اطلاعات در پزشکی

- ۱. چالش یکپارچه سازی و استخراج داده ها

داده های مراقبت های بهداشتی از منابع با فرمت های مختلف مانند داده های ساختاریافته، کاغذ، فیلم، چند رسانه ای، تصاویر دیجیتال و غیره به دست می آیند. که یکپارچه‌سازی و استخراج داده‌ها را یک چالش واقعی می‌سازد.

- ۲. داده های دائما در حال تغییر:

بیماران و پزشکان ممکن است نقل مکان کنند، نام و حرفه خود را تغییر دهند، بازنشسته شوند. سازمان‌ها همچنین جابه‌جا می‌شوند، داروها و درمان جدید معرفی می شوند.

مشکلات جمع آوری اطلاعات در پزشکی

- ۳. حریم خصوصی و مقررات امنیتی: حفظ اعتماد بیماران، پایه و اساس ساخت یک سیستم پزشکی است.

- ۴. انتظارات بیمار: صنعت بهداشت و درمان باید درک درستی از نیازهای متغیر بیماران و ترجیحات آنها داشته باشد و سپس راه حل هایی همسو با روش زندگی آنها ارائه شوند.

- ۵. . عدم وجود فرآیندهای تضمین کیفیت: چنین فرآیندی معمولا وجود ندارد. کیفیت داده ها اغلب به شخصی بستگی دارد که داده ها را وارد می نماید

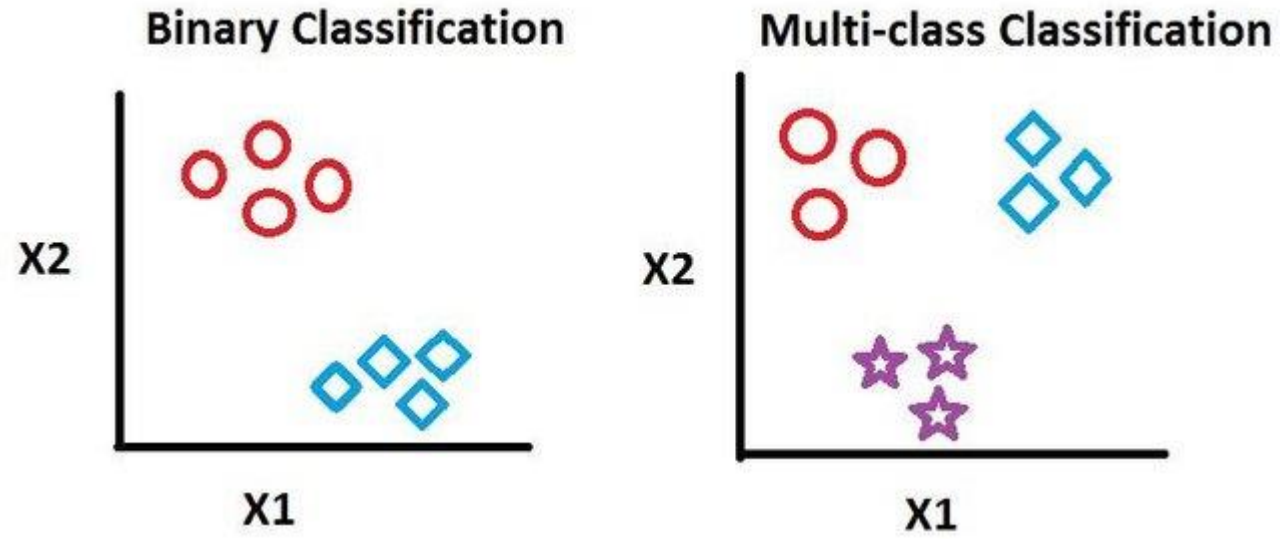
پیش پردازش داده

- انتخاب کلاس های هدف
- تکنیک‌های پاکسازی داده یا Data cleaning
- تکنیک‌های یکپارچگی داده یا Data integration
- تکنیک‌های کاهش داده یا Data reduction
- تکنیک‌های تبدیل داده یا Data transformations

انتخاب کلاس های هدف

Binary Class

Multi Class



پاکسازی داده

مدیریت داده های از دست رفته

مدیریت داده های پرت

پاکسازی داده - مدیریت داده های از دست رفته

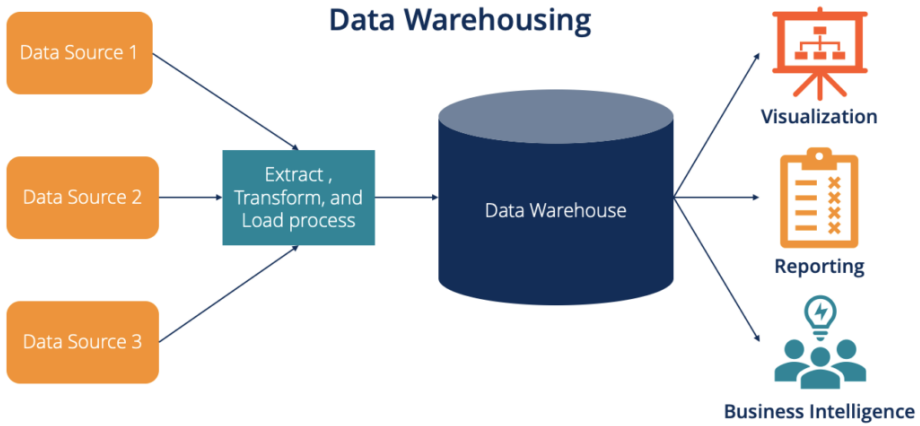
حذف یک ردیف خاص - در این روش، یک ردیفی حذف می شود که بیش از ۷۵ درصد از مقادیر آن، از دست رفته است

محاسبه میانگین - این روش برای ویژگی‌هایی که داده‌های عددی دارند مانند سن، حقوق، سال و غیره مفید است. در اینجا می‌توانید میانگین و میانه یک ستون حاوی یک مقدار از دست رفته شده است، محاسبه و جایگزین نمایید.

یکپارچگی داده

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.

استفاده از انباره داده توصیه می شود



تبدیل داده

فعالیت های مانند نرمال سازی داده ها و گسسته سازی داده

نرمال سازی داده ها : در مواقعی که فاصله ی اعداد زیاد باشد می توان با نرمال سازی محدوده ی کوتاه تری را انتخاب نمود نرمال سازی سن به بازه ی صفر تا یک

گسسته سازی داده ها : مثلا تبدیل مقادیر عددی فشارخون به سه سطح نرمال، پایین، بالا

کاهش ابعاد داده

در بسیاری از موارد، همه ی داده ها مدنظر نیستند در قسمت می توان از روش های انتخاب ویژگی استفاده نمود

مشخصات یک مجموعه ی داده ای برای استفاده در داده کاوی

۱) داشتن حداقل تعداد رکوردهای لازم

۲) داشتن حداقل تعداد صفات

۳) میزان missing value خیلی زیاد نباشد

۴) صفت هدف تا حد امکان missing value نداشته باشد

مشکلات داده کاوی در حوزه سلامت

۱) ناکامل بودن داده ها

۲) حریم شخصی

۳) کاغذی بودن اغلب پرونده ها

انواع داده

- عددی Numerical
- ترتیبی Ordinal
- اسمی Nominal

داده کاوی بر حسب نوع داده

1) متن کاوی:

متن کاوی به فرایند تحلیل و اکتشاف انبوهی از متون غیرساخت‌یافته بوسیله نرم‌افزار به منظور شناسایی مفاهیم، الگوها،
موضوعات، کلیدواژه‌ها و دیگر ویژگی‌های داده‌های متنی گفته می‌شود.

به عبارت دیگر هدف متن کاوی، کشف معنا (مفهوم و هدف) و استخراج اطلاعات نهفته (برای مثال موجودیت‌ها و روابط) در داده‌های متنی است.

داده کاوی بر حسب نوع داده

متن کاوی : کاربردها

- کسب‌وکارها : کاربردهایی از قبیل تحلیل حس و میزان رضایتمندی مشتریان نسبت به محصولات یا شرکت‌ها، گرایش و علاقه بازار نسبت به ویژگی‌های مختلف محصولات، شناسایی سلیقه یا رویدادهای زندگی کاربر و تبلیغات موثر، شناسایی خودکار و فیلتر نظرات نامناسب (غیرقابل انتشار) کاربران و ...
- کتابخانه‌ای : نمایه‌زنی و دسته‌بندی موضوعی مقالات و کتاب‌ها، مشابهت‌یابی بین مستندات مختلف، جستجوی (غیردقیق) متن یا عبارت در بین حجم انبوه منابع و ...
- جامعه‌شناسی و روان‌شناسی : تحلیل علائق، خصوصیات و خلقیات افراد، شناسایی و تحلیل لحن و نحوه بیان نشریات و رسانه‌های مختلف برای القای مقصود خود به افراد و ...
- زیست‌شناسی : نام این فیلد به متن کاوی داده‌های زیستی (Biomedical Text Mining) معروف است که بیشتر روی تحلیل تعاملات بین توالی پروتئین آنها و ارتباط و وابستگی بین آنها با بیماری‌ها با استفاده از تکنیک‌های متن کاوی تمرکز دارد.

داده کاوی بر حسب نوع داده

۲) وب کاوی

وب کاوی به دستیابی الگوهای اطلاعاتی از داده های موجود در وب کمک می کند.

بهتر شدن قدرت موتور های جستجوی وب : از طریق شناساندن صفحات وب و طبقه بندی مستندات

وب کاوی برای فهمیدن رفتار مشتری، ارزیابی اثربخشی یک وبسایت خاص مفید است

کاویدن اطلاعات مفید از صفحات وب

داده کاوی بر حسب نوع داده

3) Image mining

استخراج اطلاعات مفید از عکس ها

کووید۱۹ : استخراج اطلاعات از سی تی اسکن بیمار

داده کاوی بر حسب نوع داده

۴) داده ی جدولی (داده کاوی):

ستون ها : صفات سطرها: اطلاعات بیماران به صورت رکورد

سیستم های ثبت موجود در سطح دانشگاه و سطح کشور

انواع تکنیک های داده کاوی

۱) قواعد انجمنی :

تولید قواعد از الگوهای پرتکرار

۲) خوشه بندی :

تشخیص مجموعه ای از دسته ها برای توصیف داده مثلا تشخیص گروه های مختلف در یک شبکه

یادگیری بدون نظارت

۳) کلاسه بندی :

قرار دادن داده ها در یک دسته بندی از پیش تعریف شده

یادگیری با نظارت

انواع تکنیک های داده کاوی - قواعد انجمنی

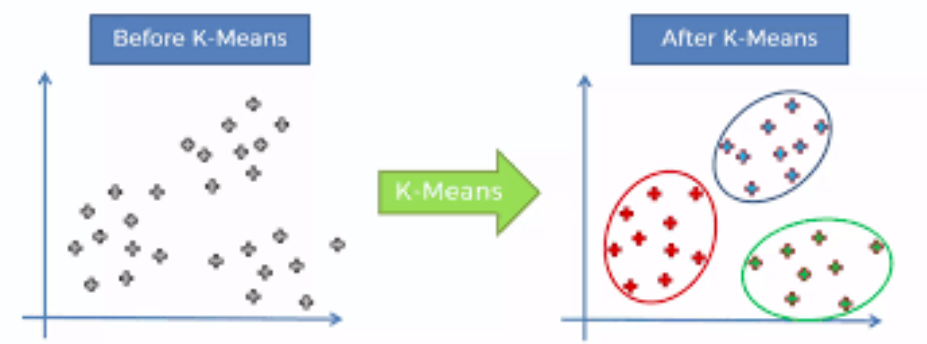
Basket	Product 1	Product 2	Product 3
1	Milk	Cheese	
2	Milk	Apples	Cheese
3	Apples	Banana	
4	Milk	Cheese	
5	Apples	Banana	
6	Milk	Cheese	Banana
7	Milk	Cheese	
8	Cheese	Banana	
9	Cheese	Milk	

- Itemset

{Milk,Cheese,Apples,Banana}

- k-itemset : زیرمجموعه ی k عنصری از Itemset
- فرض کنید زیر مجموعه ی مورد نظر : {Milk,Cheese}
- فرکانس تکرار : در شش سبد خرید Milk,Cheese با هم رخ داده اند پس فرکانس تکرار آنها برابر با ۶ است (absolutesupport)
- فرکانس تکرار نسبی : از ۹ سبد خرید، در ۶ سبد خرید Milk,Cheese با هم رخ داده اند پس فرکانس تکرار نسبی برابر با ۶۶ درصد است (relativesupport)
- آستانه کمترین تکرار : برابر با ۵۰ درصد است (min_sup)
- پس Milk,Cheese یک مجموعه ی پرتکرار است

انواع تکنیک های داده کاوی - خوشه بندی



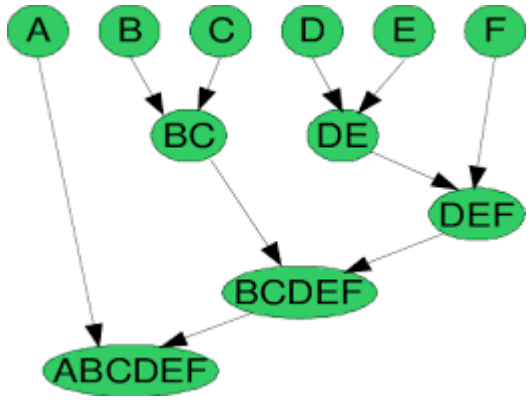
- خوشه بندی k-means

۱) k نقطه را به عنوان مرکز خوشه در نظر گرفته می شود

۲) هر شی را به شبیه ترین مرکز خوشه اختصاص دهید

۳) به مرحله 1 بازگردید ، وقتی تغییری ایجاد نشد، متوقف شوید.

انواع تکنیک های داده کاوی - خوشه بندی



- خوشه بندی سلسله مراتبی

۱. Top-down

۲. Bottom-up

انواع تکنیک های داده کاوی کلاسه بندی

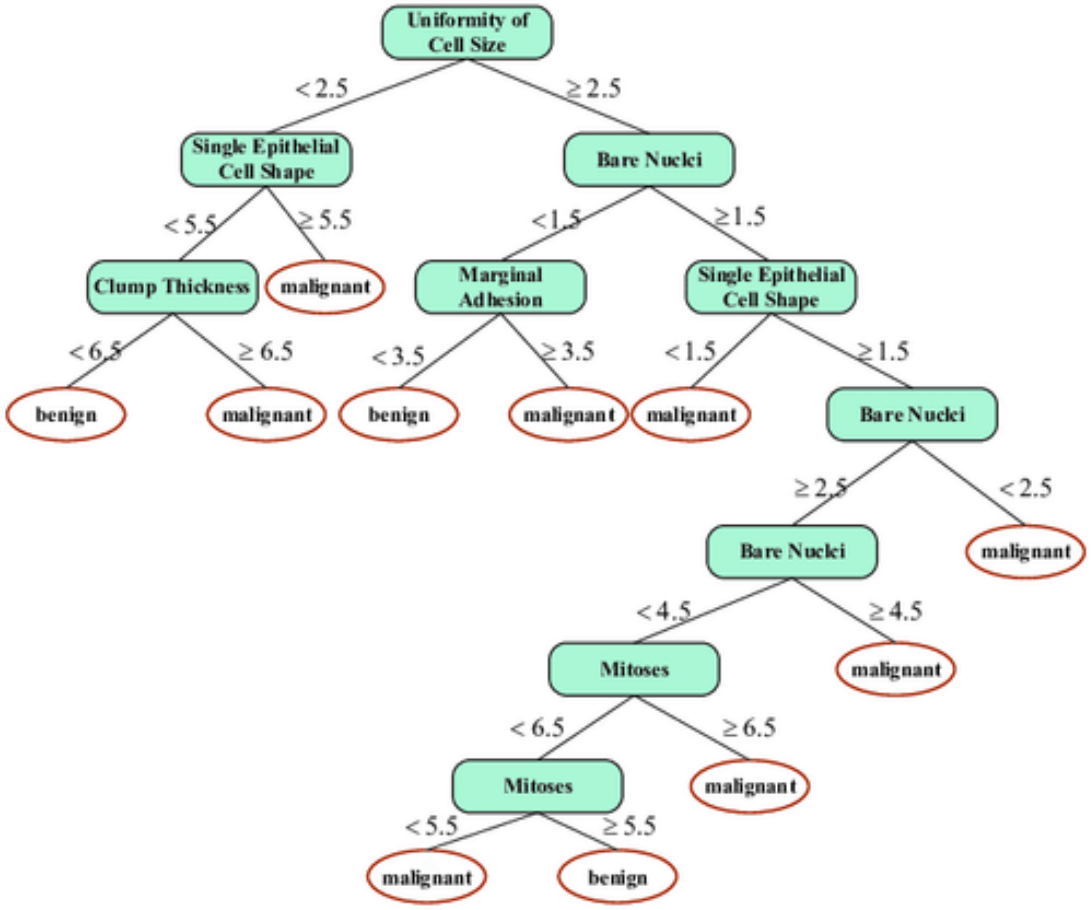
درخت تصمیم

بیزین

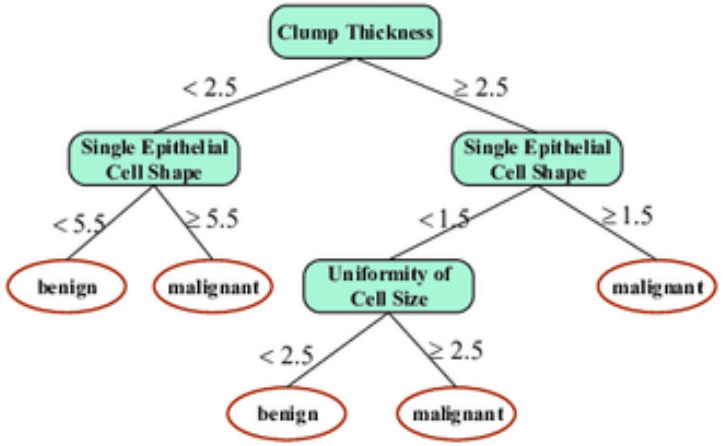
شبکه عصبی

ماشین بردار پشتیبان

درخت تصمیم برای خوش خیم و بدخیم بودن سرطان

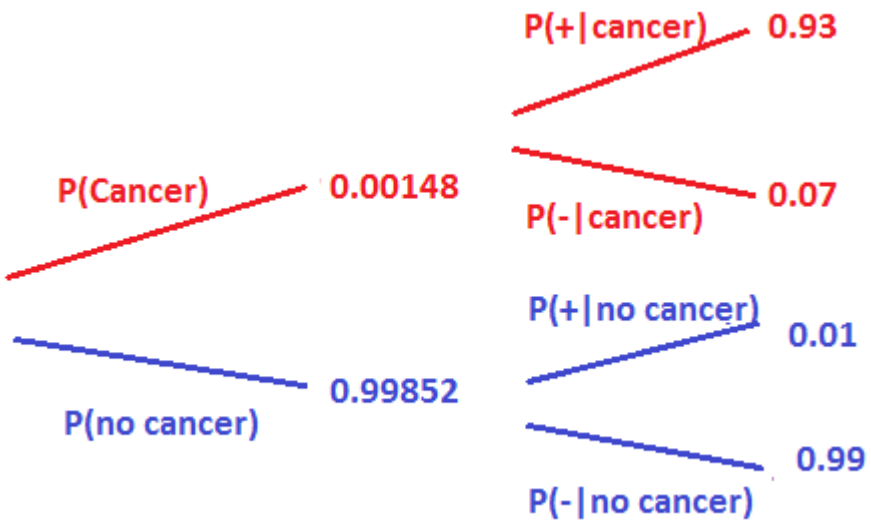


(a)



(b)

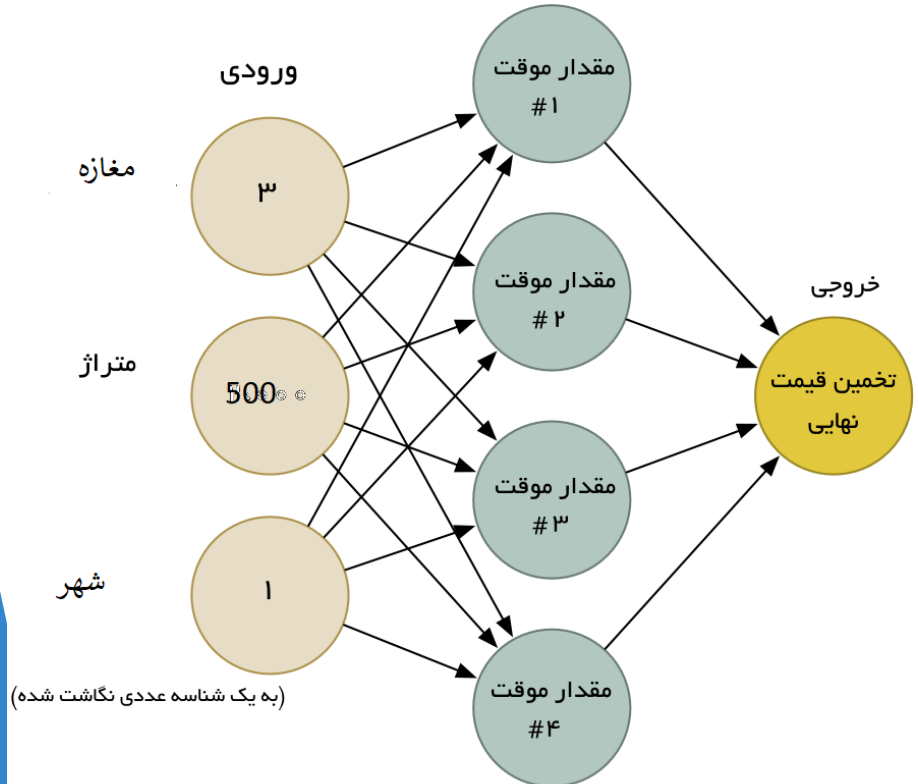
بیزین


P(+|cancer)
0.93
P(Cancer)
0.00148
P(-|cancer)
0.07
P(+|no cancer)
0.01
0.99852
P(no cancer)
0.99
P(-|no cancer)

شبکه عصبی مصنوعی

مغز انسان از نورون ها تشکیل شده که قادر هستند محاسبات معینی را سریعتر از کامپیوترهای دیجیتال انجام دهند

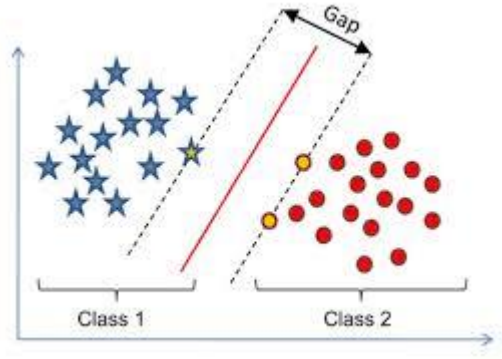
شبکه عصبی مصنوعی می خواهد عملکرد مغز را شبیه سازی نماید



ماشین بردار پشتیبان

بردارهای پشتیبان به زبان ساده، مجموعه ای از نقاط در فضای n بعدی داده ها هستند که مرز دسته ها را مشخص می کنند و مرزبندی و دسته بندی داده ها براساس آنها انجام می شود و با جابجایی یکی از آنها، خروجی دسته بندی ممکن است تغییر کند.

هدف ماشین بردار پشتیبان، بهینه کردن خط عمود بر فاصله ی بین دو کلاس است



ارزیابی تکنیک ها – 10-fold cross validation

قسمت اول، مجموعه‌ی داده‌ای آموزشی است که ۹۰ درصد مجموعه‌ی داده‌ای اصلی را تشکیل می‌دهد

قسمت دوم، مجموعه‌ی داده‌ای آزمایشی است که ۱۰ درصد مجموعه‌ی داده‌ای اصلی را تشکیل می‌دهد.

تقسیم به این دو مجموعه، ۱۰ بار انجام می‌شود و هر بار یک گروه به صورت تصادفی به عنوان مجموعه‌ی آزمایشی و بقیه‌ی گروه ها به عنوان مجموعه‌ی آموزشی در نظر گرفته می‌شوند. از مجموعه‌ی داده‌ای آموزشی برای تولید الگوها و از مجموعه‌ی داده‌ای آزمایشی برای ارزیابی صحت تکنیک‌های مورد استفاده در نظر گرفته می‌شود.

پارامترهای ارزیابی عملکرد یک کلاسه بند - تعاریف براساس کلاسه بندی دودویی

Accuracy

نسبت نمونه‌های درست پیش بینی شده به کل نمونه‌های پیش بینی شده

Sensitivity

حساسیت نسبتی از نمونه‌های مثبت است که مدل آن‌ها را به درستی به عنوان نمونه مثبت تشخیص داده است.

Specificity

نسبتی از نمونه‌های منفی است که مدل آن‌ها را به درستی به عنوان نمونه منفی تشخیص داده است

پارامترهای ارزیابی عملکرد یک کلاسه بند - تعاریف براساس کلاسه بندی دودویی

AUC

ROC : منحنی‌های ROC منحنی‌های دو بعدی هستند که در آنها Sensitivity در محور Y و 1-Specificity در محور X رسم می‌شوند.

وقتی مقدار ناحیه‌ی تحت منحنی ROC برابر ۱ باشد، دسته‌بند می‌تواند بین تمام نقاط کلاس مثبت و منفی به طور صحیح تمایز قائل شود. اما، اگر مقدار ناحیه‌ی تحت منحنی ROC برابر ۰ باشد، دسته‌بند همه منفی‌ها را مثبت و همه مثبت‌ها را منفی پیش‌بینی می‌کند.

AUC سطح ناحیه‌ی تحت منحنی ROC را نشان می دهد.

ارزیابی الگوها توسط متخصصین بالینی

در این گام، الگوها در اختیار متخصصین بالینی قرار می گیرد و سپس با دریافت نظرات آنها، الگوهای و قواعد تایید شده گزارش می شوند.

نرم افزارهای موجود

- ۱) نرم افزار رپیدماینر

قوی ترین و ساده ترین ابزار توسط کمپانی رپیدماینر تولید شده است

- ۲) نرم افزار weka

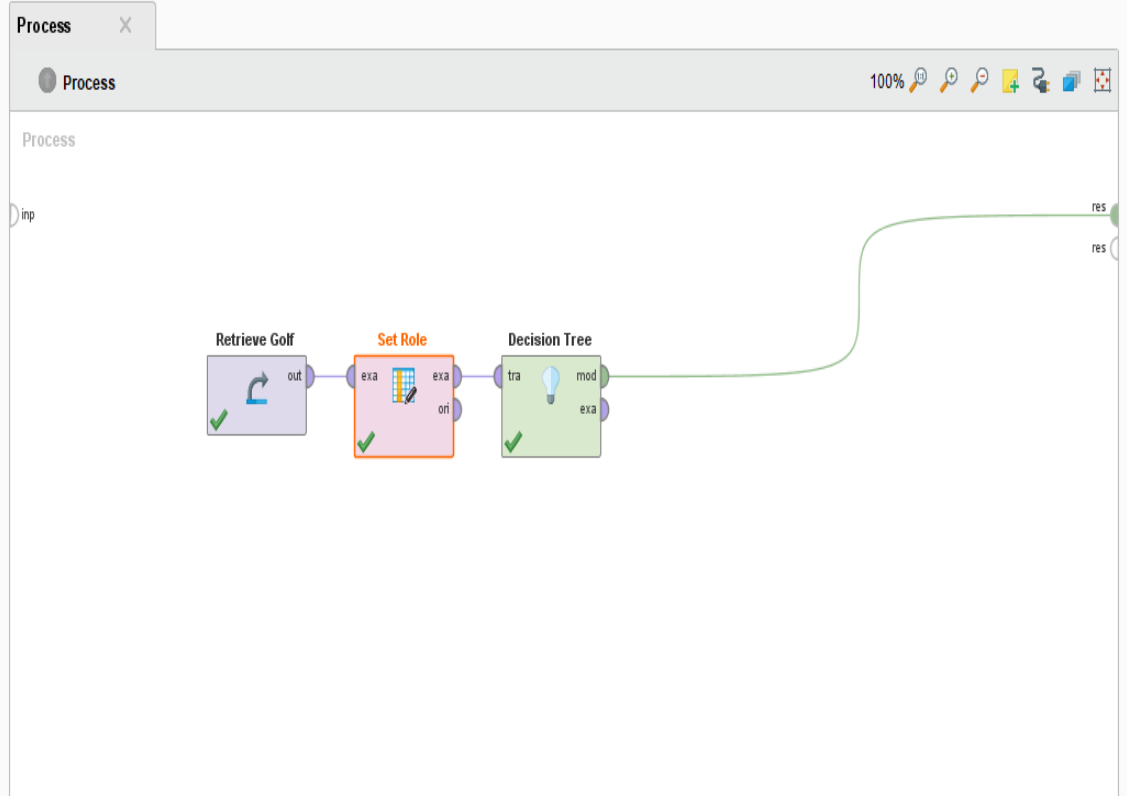
زبان جاوا، رایگان و متن باز

- ۳) Orange

به زبان پایتون

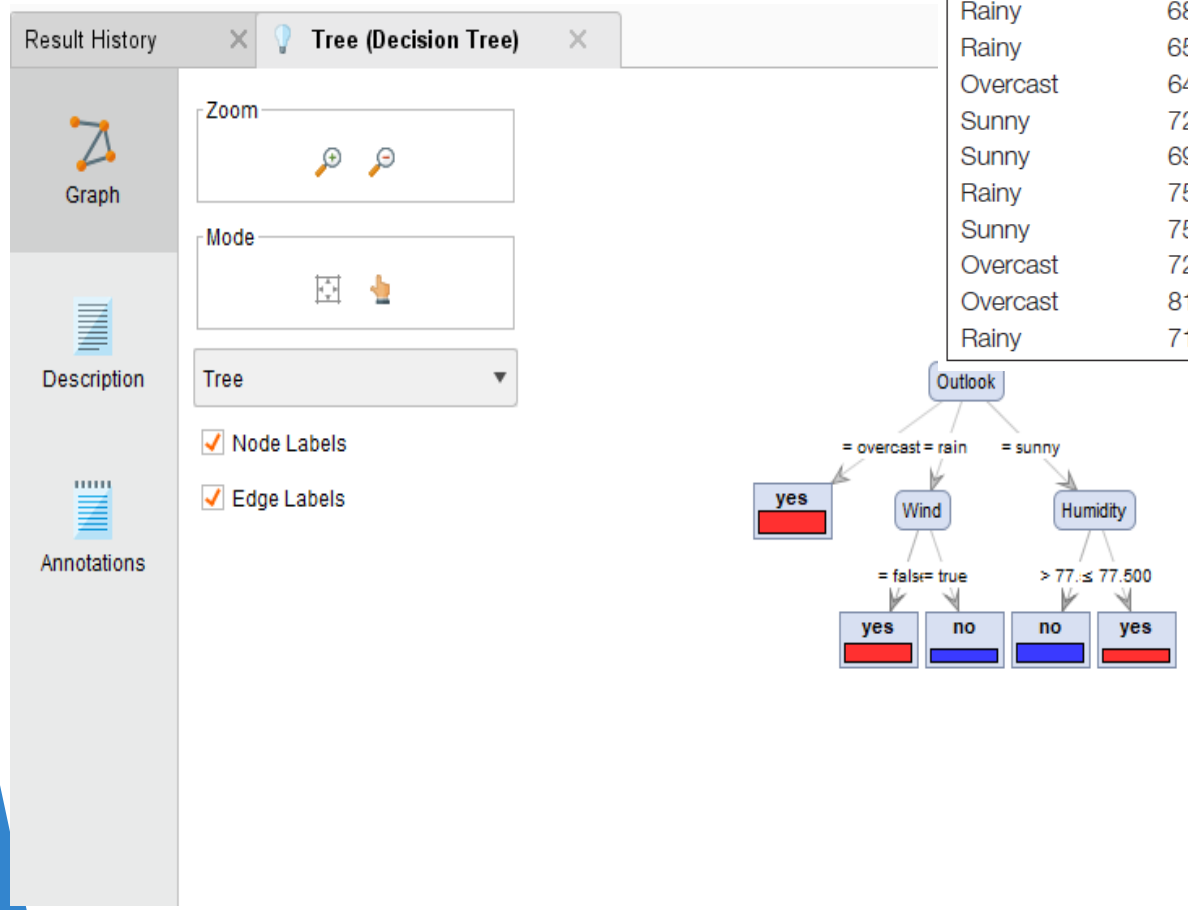
نمونه : پردازش با کلاسه بندی - رپیدماینر

انتخاب یکی از تکنیک ها به عنوان مثال درخت تصمیم



نمونه : پردازش با کلاسه بندی - رپیدماینر

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	false	no
Sunny	80	90	true	no
Overcast	83	86	false	yes
Rainy	70	96	false	yes
Rainy	68	80	false	yes
Rainy	65	70	true	no
Overcast	64	65	true	yes
Sunny	72	95	false	no
Sunny	69	70	false	yes
Rainy	75	80	false	yes
Sunny	75	70	true	yes
Overcast	72	90	true	yes
Overcast	81	75	false	yes
Rainy	71	91	true	no



حامیان و اطلاعات تماس

حامیان

معاونت بهداشتی دانشگاه علوم پزشکی اصفهان

دانشکده مدیریت و اطلاع رسانی پزشکی

مرکز تحقیقات فناوری اطلاعات در امور سلامت

گروه مدیریت و فناوری اطلاعات سلامت

ایمیل دکتر محمد ستاری msattarimng.mui@gmail.com